BerryDunn

Expert Determination Report

Expert Determination Report:
New Hampshire
Comprehensive Healthcare Information System
Public Use Data Set Redesign

Submitted to the
New Hampshire Department of Health and Human Services and the
New Hampshire Insurance Department

Submitted by:

Andrea Clark, MS
BerryDunn
Health Analytics Practice Area
254 Commercial St. 2nd Floor
Portland ME 04101
Phone: (207) 541-4900
Fax: (207) 523-8686
aclark@berrydunn.com

Submitted on:

November 16, 2017

Table of Contents

1 Executive Summary 1

2 Introduction 4

2.1 Agreed-Upon Objectives 5

2.2 Summary of Relevant Laws and Regulations 6

2.3 De-identification Standards and Guidelines 7

2.4 Compliance of Recommended Data Sets with Privacy Law and De-identification Guidelines 11

2.5 Expert Status: Qualifications and Experience 14

2.6 Outline of Succeeding Report Sections 16

3 Methodology 17

3.1 General Approach 17

3.2 Test Data 19

4 Aggregated Files 26

4.1 Medical Expense and Members Served by Payer, Provider, Service Category, and Member Region (File A1) 26

4.2 Medical Expense and Members Served by Product Type, Market Segment, and Funding Type (File A2) 33

4.3 Pharmacy Expense and Members Served by Product Type, Market Segment, and Funding Type (File A3) 40

4.4 Membership – Medical, Pharmacy, and Dental (Files A4 – A6) 46

4.5 Medical Expense and Members Served by Service and Member Demographics (File A7) 51

4.6 Medical Expense and Total Patients by Primary Diagnosis and Member Demographics (File A8) 56

4.7 Pharmacy Expense and Total Patients by Drug (File A9) 61

5 De-Identified Person-Level Files 65

5.1 Population Medical Cost by Member File (File P1) 65

5.2 Population Pharmacy Cost by Member File (File P2) 75

6 Looking forward 80

6.1 Recommended frequency of review 80

6.2 Foreseeable future challenges 80

6.3 Recommended techniques and approaches for continued compliance and analytical utility 80

7 Conclusion 82

1 Executive Summary

The increasing availability of electronic health information has facilitated the use of data for a variety of secondary uses (e.g., price transparency, epidemiological studies, or any other purpose not directly related to individuals' health care). Personal health information released for secondary use must be adequately de-identified[i] according to the standards promulgated in the November 2012 Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.[1]

The New Hampshire Comprehensive Health Care Information System (NH CHIS) was created by New Hampshire statute to make health care data "available as a resource for insurers, employers, providers, purchasers of health care, and state agencies to continuously review health care utilization, expenditures, and performance in New Hampshire and to enhance the ability of New Hampshire consumers and employers to make informed and cost-effective health care choices."[2] Since 2012, NH CHIS public use data sets have been de-identified using the HIPAA Privacy Rule Safe Harbor[ii] method and provided to the public upon written request. In an attempt to improve the utility of these data sets for secondary purposes, the New Hampshire Insurance Department (NHID) requested proposals for recommendations to redesign and enhance the public use data sets utilizing the Expert Determination[iii] standard under the HIPAA Privacy Rule.[3]

---

[i] "Reducing the risk of identifying a data subject to a very small level by applying a set of data transformation techniques such that the resulting data retains *[sic]* very high analytic utility." El Emam K, Arbuckle L. *Anonymizing Health Data: Case Studies and Methods to Get You Started.* O'Reilly Media, Inc., Sebastopol, CA 2013; p. 6.

[ii] The Safe Harbor method requires the removal of 18 direct identifiers (names, all geographic subdivisions of a state except three-digit zip code, all elements of date except year for any dates directly related to an individual or an individual's care, telephone numbers, Social Security or other individual account numbers, biometric data, etc.) and that the data custodian "not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information." United States Department of Health and Human Services Office for Civil Rights. *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.* November 26, 2012. Accessed 18 September 2017: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf: pp. 7-8.

[iii] The Expert Determination method requires that "[a] person with appropriate knowledge and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable: (i) Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and (ii) Documents the

After a competitive bid process, NHID and the Department of Health and Human Services (DHHS) contracted with the BerryDunn Health Analytics Practice Area (BerryDunn, formerly Compass Health Analytics, Inc.) to provide recommendations for redesign of New Hampshire's de-identified public use data sets using statistical methods to increase data utility while maintaining data privacy. BerryDunn staff performed literature searches, researched other states' experiences with the expert determination method, and developed a thorough understanding of the NH CHIS and data characteristics unique to New Hampshire. Informed by decades of statistics and healthcare analytics and data management experience, and in collaboration with NHID and DHHS, BerryDunn has developed the following proposed NH CHIS public use files enhancement design:

- Nine aggregated (i.e., not personal health data) data files:
  - A summary of medical expense by carrier, provider, provider city, provider state, service type, member residence in New Hampshire or other state, and member region of residence within New Hampshire for New Hampshire residents (Enhanced Public Use File A1).
  - A summary of medical expense by carrier, primary eligibility status, funding type (self-funded or fully commercially insured), product type, market segment, utilization category, member gender, member age group, member residence in New Hampshire or other state, and member region of residence within New Hampshire for New Hampshire residents (File A2).
  - A summary of pharmacy expense by carrier, primary eligibility status, funding type (self-funded or fully commercially insured), product type, market segment, drug therapeutic class, member gender, member age group, member residence in New Hampshire or other state, and member region of residence within New Hampshire for New Hampshire residents (File A3).
  - Three summary files of member months (medical, pharmacy, and dental) by carrier, primary eligibility status, funding type (self-funded or fully commercially insured), product type, market segment, member gender, member age group, member residence in New Hampshire or other state, and member region of residence within New Hampshire for New Hampshire residents (Files A4 – A6).

methods and results of the analysis that justify such determination[.]" United States Department of Health and Human Services Office for Civil Rights. *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.* November 26, 2012. Accessed 18 September 2017: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf: pp. 7.

  - A summary of medical expense by member gender, member age group, member residence in New Hampshire or other state, member region of residence within New Hampshire for New Hampshire residents, service code, and service type (File A7).
  - A summary of medical expense by member gender, member age group, member residence in New Hampshire or other state, member region of residence within New Hampshire for New Hampshire residents, first three digits of principal diagnosis reported on the claim line, and ICD10 chapter (File A8).
  - A summary of pharmacy expense by member gender, member age group, member residence in New Hampshire or other state, member region of residence within New Hampshire for New Hampshire residents, drug code (NDC), drug name (as reported by the carrier), drug brand status, and drug therapeutic class (File A9).
- Two de-identified person-level files:
  - A file of medical expenses by unique service user (identified by an encrypted member identification number), member diagnostic grouping, and utilization category (File P1).
  - A file of pharmacy expense by unique service user (identified by an encrypted member identification number) and member diagnostic grouping (File P2).
- The current claim-level release of commercial medical, pharmacy, and dental claims de-identified in a previous process using the Safe Harbor method.

The body of this report includes detailed documentation of these files, the algorithms developed to create them from the 2016 Consolidated NH CHIS commercial insurer data, and the files' compliance with current legal and commonly accepted de-identification standards. In addition, the report details BerryDunn's qualifications as an expert, discusses NHID's, DHHS's, and BerryDunn's agreed-upon objectives for the public use file redesign, and provides overviews of HIPAA Privacy Rule de-identification standards and the metrics that have been developed in academic literature and practical application to de-identify data in compliance with the rule.

The final sections look forward to likely future challenges and provide guidance for maintaining the de-identification compliance and analytical utility of the design.

2 Introduction

The New Hampshire Comprehensive Health Care Information System (NH CHIS) is New Hampshire's all payer claims database (APCD). By statute, the New Hampshire Insurance Department (NHID) and Department of Health and Human Services (DHHS) are required to collaborate in its development. DHHS is responsible for adopting rules pertaining to the release of data sets, and is required to do so to the extent allowed by the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule and state law.

The 2012 HIPAA Privacy Rule the United States Department of Health and Human Services Office for Civil Rights (OCR) offers two methods for de-identifying Protected Health Information (PHI):[iv] the Expert Determination[v] method (§164.514(b)(1)) and the Safe Harbor[vi] method (§164.514(b)(2)).[4]

[iv] Please see section 2.2 for the HIPAA Privacy Rule's definition of PHI.

[v] The Expert Determination method requires that "[a] person with appropriate knowledge and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable: (i) Applying such principles and methods, determines that the risk is very small that the information could be used, alone or in combination with other reasonably available information, by an anticipated recipient to identify an individual who is a subject of the information; and (ii) Documents the methods and results of the analysis that justify such determination[.]" United States Department of Health and Human Services Office for Civil Rights. *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.* November 26, 2012. Accessed 18 September 2017: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf: pp. 7.

[vi] The Safe Harbor method requires the removal of 18 direct identifiers (names, all geographic subdivisions of a state except three-digit zip code, all elements of date except year for any dates directly related to an individual or an individual's care, telephone numbers, Social Security or other individual account numbers, biometric data, etc.) and that the data custodian "not have actual knowledge that the information could be used alone or in combination with other information to identify an individual who is a subject of the information." United States Department of Health and Human Services Office for Civil Rights. *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.* November 26, 2012. Accessed 18 September 2017: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf: pp. 7-8.

Current release of NH CHIS public use data is via record-level claim data that incorporate the HIPAA Safe Harbor method, including additional restrictions. NHID and DHHS believed the public use data could be made more robust and consistent with the NH CHIS statute, RSA 420-G:11-a.[vii]

The departments therefore retained BerryDunn's Health Analytics Practice Area[viii] (BerryDunn) to develop a design that would improve the utility of the NH CHIS public use data sets while ensuring patient privacy by meeting the privacy standards laid out in OCR's November 2012 guidance on the Expert Determination standard.[5]

Neither the Safe Harbor nor the Expert Determination method ensures zero risk of re-identification, but BerryDunn is a strong proponent of the Expert Determination method for its flexibility and robustness to meet a variety of analytical needs while protecting patient privacy. It allows tailored solutions to the problem of balancing those considerations by employing a variety of methodologies including data suppression and generalization, encouraging the application of empirical expertise regarding the actual risk of identification, and acknowledging that different entities or data sources may require different levels of de-identification owing to varying constructive identification (i.e., re-identification from indirectly identifying attributes) risk.

2.1 Agreed-Upon Objectives

New Hampshire has been a pioneer in health care APCDs, particularly in public access to the data. It is the only state to provide access to claim-level APCD data (both public use and limited data sets) for no charge.

However, this transparency presents trade-offs. Because public use files are released to all who request them (no data use agreement or review process is required; review and approval are required for limited data set requests), the current claim-level public use files were developed using a conservative "Safe Harbor Plus" approach to reduce the risk of constructive identification. In addition to the 18 data elements proscribed by the Safe Harbor method,[ix] including all direct

---

[vii] The statute states that the CHIS "data shall be available as a resource for insurers, employers, providers, purchasers of health care, and state agencies to continuously review health care utilization, expenditures, and performance in New Hampshire and to enhance the ability of New Hampshire consumers and employers to make informed and cost-effective health care choices."

[viii] Formerly, and at the time of the contract award, Compass Health Analytics, Inc.

[ix] As discussed in Section 2.4, BerryDunn determined that the current and proposed public use file designs collectively pose a very small re-identification risk, in compliance with the expert determination method. While BerryDunn was not engaged to assess the compliance of the current public use files with the Safe Harbor method, we note the current files include member county of residence. The Safe Harbor method as described in the HIPAA Privacy Rule bars all geographic elements below the state level except three-digit zip

member identifiers, no payer-identifying data elements are included in the public use files, and provider information is limited.

This project seeks to enhance the data available in the face of these trade-offs, continuing New Hampshire's commitment to health care market data transparency and accessibility, not only for health care researchers, but also engaged citizens. The departments retained BerryDunn to develop the new design with the stated goal of creating public use data offerings that do the most good for the greatest number (of data users and other stakeholders) within the constraints of the HIPAA Privacy Rule.

2.2 Summary of Relevant Laws and Regulations

The NH CHIS statute, RSA 420-G:11-a., states that the "data shall be available as a resource for insurers, employers, providers, purchasers of health care, and state agencies to continuously review health care utilization, expenditures, and performance in New Hampshire and to enhance the ability of New Hampshire consumers and employers to make informed and cost-effective health care choices." As noted, by statute, NHID and DHHS are required to collaborate in developing NH CHIS, and DHHS is responsible for adopting rules pertaining to the release of data sets.

The HIPAA Privacy Rule requires PHI to be de-identified before release for secondary use.[x] That is, information that could directly identify individuals in the data (names, account numbers, addresses, etc.) must be masked or removed, and information released that could indirectly identify individuals (birthdates, diagnoses, procedures, etc.) must be determined to be very unlikely to result in the successful re-identification of individuals.[6]

OCR provides the following definition of PHI in the final HIPAA Privacy Rule:

> "The HIPAA Privacy Rule protects most "individually identifiable health information" held or transmitted by a covered entity or its business associate, in any form or medium, whether electronic, on paper, or oral. The Privacy Rule calls this information protected health information (PHI). Protected health information is information, including demographic information, which relates to:

code, and requires that three-digit zip code must be masked (that is, reported as '000') if the zip code has fewer than 20,000 residents. Exactly one of New Hampshire's three-digit zip code areas has fewer than 20,000 residents; reporting that single three-digit zip code as '000' would do nothing to de-identify that area. Since all New Hampshire counties have more than 20,000 residents, using county as the basis of geographic sub-division prevents a geographic unit with fewer than 20,000 residents from being identified.

[x] Any use not directly related to an individual's medical care or health care insurance coverage, such as public health or social sciences research, and for which an individual's consent has not been obtained.

> the individual's past, present, or future physical or mental health or condition,
>
> the provision of health care to the individual, or
>
> the past, present, or future payment for the provision of health care to the individual
>
> and that identifies the individual or for which there is a reasonable basis to believe can be used to identify the individual."[7]

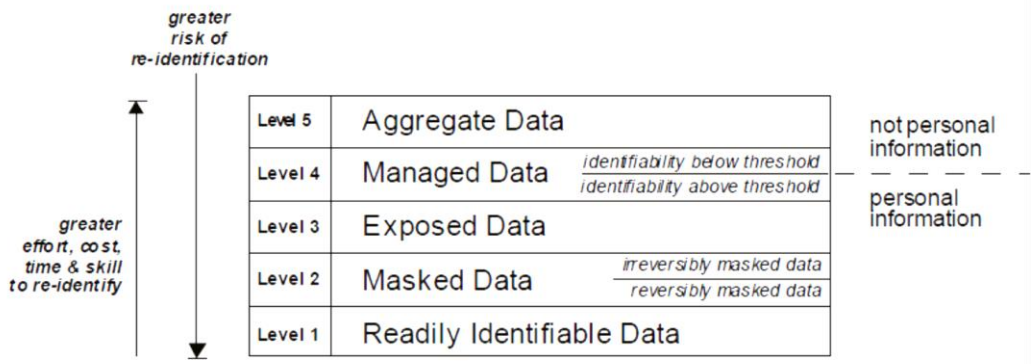
The HIPAA Privacy Rule was designed to protect individuals by limiting unauthorized uses and disclosures of individuals' Protected Health Information (PHI). It offers two methods for de-identifying PHI: the Expert Determination method (§164.514(b)(1)) and the Safe Harbor method (§164.514(b)(2)).[8] While neither method ensures zero risk of re-identification, both facilitate secondary use of data for comparative effectiveness studies, policy assessment, life sciences research, and other endeavors.[9] The Safe Harbor method is often used because HIPAA defines it more precisely, and it is less expensive and arduous to implement. Expert Determination, in contrast, relies upon the application of statistical or scientific principles that result in only a very small risk that an individual could be identified.

The Privacy Rule requires the risk of re-identification to be "very small" but sets no explicit numerical threshold. BerryDunn worked with NHID and DHHS to determine the level of risk the agencies deem acceptable and to confirm agreement on the proposed solution. Some risk of re-identification will necessarily remain, and indeed is present even with the Safe Harbor method.

2.3 De-identification Standards and Guidelines

A robust academic literature provides standards and guidelines for understanding re-identification risk and implementing de-identification using statistical methods (the Expert Determination method is also referred to as the "statistical method" of HIPAA de-identification[10]). Figure 1 provides a useful visualization of types of data and their corresponding levels of identification risk.

Figure 1
Types of Data and Levels of Identifiability



Source: El Emam, K. "Method and Experiences of Risk-Based De-identification." CHEO Research Institute, 2010. http://www.ehealthinformation.ca/media/events/webinars/de-identification-case-studies/. Accessed 31 July 2017.

Health care claim data as maintained by health insurance carriers, which by necessity include unmasked direct personal identifiers (so carriers can communicate with their members about their coverage) are an example of Level 1 data as defined by Figure 1. The goal of the Expert Determination process when preparing data for public use is output comprising Level 4 "identifiability below threshold" and/or Level 5 data, that is, data sources that no longer comprise (identifiable) personal information, or PHI (although Level 4 data may be presented at the person level).

Re-identification risk management can be implemented via a variety of techniques, including:

- *Aggregation:* Summarizing the data before releasing it, such that any row or field shows data for multiple individuals.
- *Masking:* The process of removal (suppression), randomization, or recoding of direct identifiers. All direct identifiers in a data set must be removed or masked in order for person-level data to be considered de-identified.[11]
- *Generalization:* "Reducing the precision of a field[,]"[12] such as by changing age on date of service to a ten-year age interval.
- *Suppression:* Removing a row or column of data or replacing it with a null value,[13] [14] this technique is required where adequate cell size (in the case of aggregate data) or equivalence classes cannot be achieved via the other common de-identification methods.
- *Data Truncation:* Also known as *top-coding,*[15] in this technique, "values that are considered uncommonly high" and may therefore pose a re-identification risk are cut off and an upper bound value is reported instead.[16] This technique can be used to create a lower bound as well.
- *Sampling:* Releasing only a random sample of the source data.[17]

- *Perturbation:* Changing the values of data in some fields, also known as "adding noise."[18]

A single successful de-identification effort will typically use multiple techniques. BerryDunn's efforts on behalf of DHHS and NHID have primarily used aggregation, generalization, truncation, and suppression.

The following sub-sections describe public-release standards for aggregated and person-level data.

Aggregated Data

Aggregated data are not personal health information, and, as such, do not require de-identification risk management; they are therefore adequately de-identified by definition. However, care must be taken in presenting aggregated data for public use that the data presented are, in fact, aggregate data, and cannot be re-formatted or manipulated to expose individual health information. This translates to requiring that a minimum amount of data be summed together in any given cell or record on a table in order for the data to be considered aggregated, not personal, data. For example, a table counting distinct individuals with a given principal diagnosis that reported one patient with a particular diagnosis would effectively comprise person-level, not aggregated data, so any given record in such a diagnosis table must report multiple people with the same diagnosis to be considered aggregate data. This requirement is often called a *minimum cell size standard*.

A range of minimum cell sizes for aggregated tables have been promulgated in practice and in the de-identification literature, with published recommendations ranging at least from three to twenty, and the most common threshold in practice being five.[19]

In the present case, BerryDunn followed the United States Centers for Medicare and Medicaid Services' (CMS) cell suppression guidelines requiring a minimum cell size of eleven, that is, no result with a value of one to ten (for measures such as admissions, services, users, etc.) may be reported.[20] Zeroes, however, are reportable.

Person-Level Data

Person-level files comprise personal health information until re-identification risk has been adequately managed; for public release data, managing re-identification probability risk to a maximum of five percent for a data set's most vulnerable record is a fairly conservative and commonly-accepted standard according to the academic de-identification literature.[21] Vulnerability is generally understood as a function of information directly identifying an individual (names, individual identifying numbers, etc.) and uniqueness of indirectly identifying information (birthdates, diagnoses, etc.). Therefore, re-identification risk can be reduced to or below a given threshold by removing or encrypting directly identifying information and removing uniqueness.

This standard requires encryption of any *direct identifiers,* those pieces of information, such as names or social security numbers, that directly identify a person or entity, and the creation of

adequately large *equivalence classes.* An equivalence class is a set of records that shares "the same values on a set of quasi-identifiers."[22]

Quasi-identifiers, or indirectly identifying variables, are data elements that:

- Do not in and of themselves uniquely identify an individual (e.g., birthdate),
- May be used by a data user with background information on the data element "individually or in combination, to re-identify a record with high probability,"[23] and
- Are useful for data analysis.[24]

A simple metric for assessing re-identification risk, or the probability of re-identification (Pr(ID)), posits that re-identification risk is no greater than the inverse of the size of the minimum equivalence class. That is, given that a probability of re-identification less than or equal to five percent is equivalent to:

$$\Pr(\mathrm{ID}) \leq 1/20$$

requiring equivalence classes of at least 20 for all combinations of quasi-identifiers and effectively encrypting direct identifiers meets the generally-accepted 5% re-identification risk standard of de-identification of person-level health care data for public use. However, there are additional sources of re-identification risk that must be considered.

After masking direct identifiers and building adequate equivalence classes to manage re-identification risk within a data set, the data custodian must consider the risks of identification of subjects in the data set through linkage to other data sources (e.g., voter registration, vital records databases, or other tables or exhibits derived from the same source data). Releasing a random sample of the data is a powerful tool against this risk, as it creates uncertainty whether or not any given individual is in the data set, and (as follows intuitively), the probability that an adversary will be successful in attempting to re-identify a particular individual is always lower if the adversary does not (cannot) know the person is in the data set than if she does.

Sampling has a second, practical advantage for the developer of the de-identification algorithm: when a sample of personal data are taken prior to de-identification, the equivalence class size required for de-identification can be reduced in proportion to the size of the sample.[25] That is, if adequate de-identification of a population data set requires equivalence classes of at least 20, a 50% random sample of the population data set will require equivalence classes of 10 to meet the same de-identification standard.

Finally, although quantitative data elements (allowed charges, paid dollars, claim counts, drug days supply, etc.) are neither direct identifiers nor quasi-identifiers, extreme values reported in these fields are potential additional sources of uniqueness or near-uniqueness, and in the most extreme cases may be identifying (for example, a single case of a new, very expensive treatment being tried for a certain medical condition could make the news, providing an adversary adequate background

knowledge to attempt re-identification of the patient's data). Therefore, truncation of extreme values, or top-coding, is typically indicated when de-identifying person-level data.

2.4 Compliance of Recommended Data Sets with Privacy Law and De-identification Guidelines

The HIPAA Privacy Rule was designed to protect individuals by limiting unauthorized uses and disclosures of individuals' protected health information. As noted above, BerryDunn is a strong proponent of the Expert Determination, or statistical, method of de-identification under the HIPAA Privacy Rule for its flexibility and robustness to meet a variety of analytical needs while protecting patient privacy.

The Expert Determination method balances analytical utility (or increased utility compared with that available with the Safe Harbor method) with the risk of re-identification (similarly, increased risk compared to risk with the Safe Harbor method). Incremental risk is unavoidable; that is, any loosening of the Safe Harbor standards creates some additional risk, however small, of re-identification, compared to the risk of releasing a data set meeting the Safe Harbor standard.

The Privacy Rule requires that the risk of re-identification be "very small" but sets no explicit numerical threshold. BerryDunn worked with NHID and DHHS to determine the level of risk the agencies deem acceptable and confirm agreement on the proposed solution.

The public use data enhancement proposed by BerryDunn adds eleven new data files to the currently-available claim-level medical, pharmacy, and dental public use files de-identified using the Safe Harbor method. The design includes nine aggregated data files:

- A summary of medical expense by carrier, provider, provider city, provider state, service type, member residence in New Hampshire or other state, and member region of residence within New Hampshire for New Hampshire residents (Enhanced Public Use File A1).
- A summary of medical expense by carrier, primary eligibility status, funding type (self-funded or fully commercially insured), product type, market segment, utilization category, member gender, member age group, member residence in New Hampshire or other state, and member region of residence within New Hampshire for New Hampshire residents (File A2).
- A summary of pharmacy expense by carrier, primary eligibility status, funding type (self-funded or fully commercially insured), product type, market segment, drug therapeutic class, member gender, member age group, member residence in New Hampshire or other state, and member region of residence within New Hampshire for New Hampshire residents (File A3).
- Three summary files of member months (medical, pharmacy, and dental) by carrier, primary eligibility status, funding type (self-funded or fully commercially insured), product type, market segment, member gender, member age group, member residence in New

Hampshire or other state, and member region of residence within New Hampshire for New Hampshire residents (Files A4-A6).

- A summary of medical expense by member gender, member age group, member residence in New Hampshire or other state, member region of residence within New Hampshire for New Hampshire residents, service code, and service type (File A7).
- A summary of medical expense by member gender, member age group, member residence in New Hampshire or other state, member region of residence within New Hampshire for New Hampshire residents, first three digits of principal diagnosis reported on the claim line, and ICD10 chapter (File A8).
- A summary of pharmacy expense by member gender, member age group, member residence in New Hampshire or other state, member region of residence within New Hampshire for New Hampshire residents, drug code (NDC), drug name (as reported by the carrier), drug brand status, and drug therapeutic class (File A9).

To ensure adequate aggregation, BerryDunn followed CMS cell suppression guidelines requiring that no result with a value of one to ten may be reported 1 to 10 (for measures such as admissions, services, users, etc.) may be reported.[26] The design includes a hierarchical generalization strategy for each file that successively generalizes values of the data grouping variables until cell sizes of at least 11 for the claim count, user and patient count, and days supply (pharmacy only) fields are achieved for each row in the expense files. Any dollar value is considered reportable. A threshold of 132 member months, or the number of member months representing 11 continuously-eligible individuals, is used for the member month files.

After all fields have been generalized, any remaining records with inadequate cell size in *any* of the tested fields are suppressed.

In testing, all material amounts were reportable in all nine files.

Utilizing these standards puts the nine aggregated files in the enhanced public use data set design in compliance with a significantly more conservative minimum cell size standard than the most commonly used minimum cell size standard, five.

Two person-level files complete the enhanced file design:

- A file of medical expenses by unique service user (identified by an encrypted member identification number), member diagnostic grouping, and utilization category (File P1).
- A file of pharmacy expense by unique service user (identified by an encrypted member identification number) and member diagnostic grouping (File P2).

As person-level files, these files comprise personal health information until re-identification risk has been adequately managed; for these files, re-identification risk has been managed to the five percent maximum standard discussed above. The standard requires encryption of any direct

identifiers and the creation of adequate equivalence classes on combinations of released quasi-identifiers.

Files P1 and P2 include the person-level identifier PERSON_KEY. The source field for this variable on the Consolidated NH CHIS (also called PERSON_KEY) is an encrypted member identifier algorithmically derived by the departments' data vendor from carrier submissions. The values released on files P1 and P2 have been separately re-encrypted as a safeguard against linking the data files to each other or other NH CHIS data sets. The files therefore meet the encryption of direct identifiers standard. Sections 5.1 and 5.2 discuss the re-encryption of PERSON_KEY in more detail.

The diagnostic grouping and utilization categories are the quasi-identifiers in the medical person-level cost file (File P1); diagnostic grouping is the sole quasi-identifier in the pharmacy person-level file (File P2). In addition to reducing the risk of constructive identification by making adequate equivalence classes easier to achieve (i.e., it is easier to find records with identical values reported in one or two data fields than it is to do the same for three, five, or ten fields), limiting the number of released quasi-identifiers reduces re-identification risk in and of itself by providing a potential adversary[xi] less information about the data subjects.

As discussed in the preceding section, requiring equivalence classes large enough to result in a re-identification risk of five percent or less for any reported record, when combined with the encryption of direct identifiers, satisfies a commonly-accepted standard of de-identification.

The de-identification process must also account for and ameliorate the risk of the de-identified data being linked to other data sources to create larger pools of potentially identifying information on the data subjects. Releasing random samples of de-identified data sets is a key tool for mitigating this linkage risk. Therefore, files P1 and P2 include random 50% samples of members[xii] with paid utilization, providing security against attempts to re-identify observations in the person-level files by linking to additional data in other data sources (particularly other files in the enhanced NH CHIS public use files design). The re-encryption of PERSON_KEY, and the use of separate re-encryption algorithms for files P1 and P2 (that is, the same member is randomly assigned separate re-encrypted PERSON_KEYs in the two files), provides an additional safeguard against data linkage within the NH CHIS.

As noted above, BerryDunn assumed throughout the design process that the current public use release of commercial medical, pharmacy, and dental Safe Harbor claim files would be part of the final enhanced public use file design. BerryDunn therefore considered the Safe Harbor files a source of linkage and constructive identification potential while designing the proposed enhanced files. It

[xi] The adversary is the term of art in de-identification literature for an individual or group attempting to re-identify individuals in de-identified data. This is also called conducting a re-identification attack.

[xii] As defined by the NH CHIS variable PERSON_KEY.

is possible that an Expert Determination process to create a public use file design *excluding* the Safe Harbor claim-level files could include additional quasi-identifiers and detail; assessing this with more certainty would require reformulating and repeating the adversary threat analysis.

In addition, sampling bestows a second significant advantage: when a sample of personal data are taken prior to de-identification, the equivalence class size required for de-identification can be reduced in proportion to the size of the sample.[27] In this case, the required equivalence class size for the files P1 and P2 to meet the five percent threshold can be reduced by half to 10. However, the current design uses generalization and suppression (though no suppression was required in testing) to achieve a minimum equivalence class size of 11; this slightly more conservative rule standardizes the aggregated file cell size and person-level file equivalence class size requirements in the design while achieving adequate de-identification.

Extreme values of quantitative variables, such as health care expenses or claim counts, are an additional potential source of uniqueness, leading to re-identification risk. Therefore, very high member expenses, drug days supply, and counts of claim lines are capped in files P1 and P2. Similarly, a floor of zero has been applied to any negative member-level results for these same quantitative variables.

Finally, BerryDunn subjected the final design to a re-identification vulnerability review conducted by a Health Analytics Practice Area senior analyst who had not been involved in the development of the design, but who had strong interest in the project and issues of data privacy. The senior analyst identified no areas of concern.

The data fields included in each file, the generalization, truncation, and suppression algorithms used to achieve the above standards of data security in each file, and the results (amounts requiring generalization and/or suppression at each step in the algorithm) for each file in the testing process are discussed in detail in Sections 4 and 5, below.

2.5 Expert Status: Qualifications and Experience

The HIPAA Privacy Rule provides guidance on who qualifies as an "expert" for purposes of making an Expert Determination and rendering health information de-identified, but specifies no particular academic or professional degree or certification program. However, the expert must have appropriate knowledge of and experience with generally-acceptable statistical and scientific principles and methods for rendering information not individually identifiable. Determining if someone is an expert, according to OCR, involves a review of relevant professional experience and academic or other training of the expert, as well as the actual experience of the expert using health information de-identification methodologies.

BerryDunn's expertise in the following qualifies them to assist NHID and DHHS in designing Public Use data sets:

- HIPAA and related federal and state privacy laws and regulations

- Statistical and scientific principles and methods of de-identification
- Data structure and content for APCDs, in particular the NH CHIS, including a detailed understanding of the domains (allowed values) of important fields

Qualifications in privacy law

Through work with New Hampshire, as well as other government and private sector health care clients, BerryDunn staff have stayed abreast of HIPAA rules and regulations. They have been consulting with clients on a wide range of HIPAA compliance issues, including data privacy concerns, since 2004. One of our staff is certified in healthcare privacy and security.

Qualifications in statistical and scientific principles and methods of de-identification

BerryDunn staff have formal training in statistical disciplines. Andrea Clark, who provided the majority of effort on this project, holds a master's degree in economics with a substantial statistical component, and has conducted an extensive review of the academic health care data de-identification literature. Staff members Tina Shields and Jennifer Elwood each hold a master's degree in statistics, and staff members Judy Loren and Devin Anderson each hold degrees with ancillary concentrations in statistics. Several staff have past job experience that includes roles involving statistical analysis. All staff members were available to lend their expertise to the project during the duration of the project.

Qualifications arising from real-world NH CHIS and other APCD experience

A practical understanding of what data within the source data are actually populated, valuable in statistical analysis, and potentially exploitable by someone wishing to compromise privacy, is essential to a successful de-identification methodology. BerryDunn's extensive experience with several state APCDs, and the NH CHIS in particular, positions our staff as subject matter experts qualified to develop enhanced de-identified public use data and assess re-identification risk.

BerryDunn has extensive experience working with APCDs, including the NH CHIS and the Massachusetts APCD (MA APCD). Arising from this work, and central to the project, is a deep understanding of the data structure and content of APCDs, in particular the NH CHIS, including detailed knowledge of the domains (allowed values) of important fields (e.g., birthdates, dates of specific services, address data, etc.) and the distribution and sparsity of those values.

The MA APCD data application process requires applicants to adhere to a "minimal use" standard and to justify each data field requested, requiring a field-by-field analysis of the balance between identifiability and data utility. BerryDunn has successfully navigated this process, meeting client analytical needs while maintaining data privacy within the terms of its data use agreement, four times since the preliminary release of the MA APCD.

2.6 Outline of Succeeding Report Sections

The report proceeds in the following sections:

- Section 3 provides a detailed overview of BerryDunn's public use files enhancement methodology and testing process, and a brief description of the resulting 11 files.
- Section 4 describes each of the 9 aggregated enhanced CHIS public use files, and their derivations, in detail, including statistical results of the testing process.
- Section 5 describes the 2 person-level enhanced CHIS public use files, and their derivations, in detail, including statistical results of the testing process.
- Section 6 looks to the future, with recommendations for implementation by the departments' data vendor and discussion of foreseeable future challenges.
- Section 7 provides a brief conclusion.

3 Methodology

3.1 General Approach

To design a public use data enhancement strategy for the NH CHIS, BerryDunn reviewed the following:

- The current NH CHIS public use files
- The Consolidated NH CHIS
- Academic literature on personal health information de-identification standards
- Public use APCD data offerings of other states
- Available information on de-identification methodologies used by those states

This review led to the conclusion that the application of adequate de-identification to claim-level public use data sets enhanced with Safe Harbor-excluded data elements would require a level of data generalization resulting in a profound loss of analytical utility.

BerryDunn therefore recommends supplementing the current claim-level NH CHIS Safe Harbor public use files with a set of 11 files (tables) as follows:

- An aggregated file of medical **costs and utilization by payer, provider, service category, and member residence geography** (File A1). Aggregated data do not technically constitute personal health information; however, to protect against potentially identifying unique or near-unique records, BerryDunn's recommended methodology applies CMS cell suppression standards (i.e., no value greater than 0 or less than 11 may be displayed for any measure of claims, services, units, or patients served; any dollar value is reportable) to all the proposed aggregated files.
- Two (medical and pharmacy) aggregated files of **cost and utilization by payer, product type, funding source, market segment, utilization type** (general service utilization category or drug therapeutic class), **and member demographics** (age group, gender, and geography, (Files A2 – A3). CMS cell suppression standards are applied.
- Three (medical, pharmacy, and dental) aggregated files of **member months by payer, product type, funding source, market segment, and member demographics** (Files A4 – A6). CMS cell suppression standards are applied.
- An aggregated file of **medical expense, utilization, and patients by service code** (revenue code or HCPCS/CPT4 code), **service category, and member demographics** (File A7). CMS cell suppression standards are applied.
- An aggregated file of **medical expense, utilization, and patients served by principal diagnosis** (first three digits of ICD-10 diagnosis code), **ICD-10 chapter, and member demographics** (File A8). CMS cell suppression standards are applied.

- An aggregated file of **pharmacy expense, utilization, and patients by drug**, (NDC code and drug name), **drug therapeutic class, and member demographics** (File A9). CMS cell suppression standards are applied.
- Two files (medical and pharmacy) of **cost and utilization by member** (Files P1-P2). File P1 summarizes member expense and utilization by general diagnostic category and high-level utilization category. File P2 summarizes member expense and utilization by general diagnostic category only. As member-level files, these files constitute personal health information, and require de-identification, as discussed below.

The general methodology and goals for creating each file is described below.

1. Each file in the proposed design includes information not available on the current public use files.
2. Each file includes one calendar year of data based on incurred date. BerryDunn recommends producing the files annually, when 6 months of paid claims runout becomes available (2016 production files would be run on data for services rendered January – December 2016 and paid through June 2017).
3. In testing, ten of eleven files required some degree of data generalization to achieve the required privacy standard. The generalization strategy was organized around the following goals:
    a. Minimizing required data suppression (i.e., amounts unreportable because adequate cell or equivalence class size could not be obtained).
    b. Maintaining the integrity of field domains.
    c. The order in which variables are generalized for each file proceeds such as to preserve information unique to a given file as long as possible.
        i. Demographic data are generalized first in files where they appear.
        ii. While this approach results in larger volumes of data being reported with masked demographic characteristics, it minimizes the volume of data reported with aggregated values of health care analysis-specific characteristics such as funding source (fully-insured vs. self-insured), service codes, or diagnosis codes.
4. The de-identification strategy accounts for the possibility of linkage between the proposed files, between the proposed files (separately or in combination) and existing NH CHIS files, and between the proposed files and external data sources.
5. BerryDunn recommends releasing the 11 enhanced NH CHIS public use files as comma-delimited value (.csv) text files.

3.2 Test Data

Universe of Commercial Insurance Data

Testing and design were performed on an extract from the Q4 (fourth quarter) 2016 (claims paid through December 2016) Consolidated NH CHIS for service year 2016. The current public use files include only commercial claims; accordingly, the test data set was limited to commercial data only. In addition, only paid claims and reversals were included in the testing and draft files. Denied line items were excluded.

The medical claims files were tested on a data extract from the NH CHIS Claim_MC_2016 table. The pharmacy claims files were tested on a data extract from the Claim_PC_2016 table. The medical, pharmacy, and dental membership files were tested on data extracts from the Membership_MC_2016, Membership_PC_2016, and Membership_DC_2016 tables, respectively.

In all files including a payer data element, payer was defined by the NH CHIS field PARENT_PAYER_CODE (MC001, PC001, DC001). The COMPANY field from the NH CHIS reference table Ref_Payer was used as the company name.

All data for providers not submitting claims and membership files through at least March 2016 were excluded.

Data for one payer, "NHC0752 = United Healthcare Insurance Co. – Medicare & Retirement" had medical claims in the Consolidated NH CHIS coded as commercial claims, but no medical membership coded as commercial. Two such payers, "NHC0752 = United Healthcare Insurance Co. – Medicare & Retirement" and "NHC0152 = Humana Insurance Company," existed in the pharmacy claims. Data for these payers were excluded from the test data universe.

The methodology does *not* limit reported claim data to claims for which a valid membership record can be found for the member and month of service, nor does it limit results to claims and membership for a member's primary insurance (e.g., where the member has two insurers, one of which is a behavioral health "carve-out").

Observation counts and totals for key quantitative variables for the resulting test files are as follows:

- Medical Claims
  - 13,237,837 observations (12,464,227 claim lines, less reversals)
  - $2,374,869,392 allowed
  - $1,889,306,120 paid (includes prepaid amounts)
  - 553,543 distinct users (i.e., distinct values of the NH CHIS variable PERSON_KEY)
- Pharmacy Claims

  - 10,101,515 observations (8,331,465 claim lines, less reversals)
  - $1,017,093,185 allowed
  - $941,396,349 paid
  - 611,611 distinct users (i.e., distinct values of the NH CHIS variable PERSON_KEY)
- Medical Membership
  - 6,154,773 observations and member months
- Pharmacy Membership
  - 9,263,847 observations and member months
- Dental Membership
  - 2,310,222 observations and member months

BerryDunn has provided its programming code to DHHS, NHID, and the departments' data vendor to provide detailed documentation of the subsetting methodology and facilitate initial production of the enhanced public use files.

Transformations and Value-Added Fields

BerryDunn made the following transformations and generalizations in the test data universe for all files:

- Positive dollar or drug days supply amounts were multiplied by -1 for reversal (SV_STAT = "R") records.
- A claim line counter was added to each record. The counter was set equal to 1 for paid (SV_STAT = "P") records, and to -1 for reversal (SV_STAT = "R") records.
- As noted above, payers with claims but no commercial member months reported in eligibility files were deleted from claims. This affected one payer in medical claims and two in pharmacy.
- Any medical or pharmacy payer code that does not account for at least 1% of paid claim dollar volume was aggregated into a new payer code, "OTHPAYR = Other or Unknown Payer."
- Service codes with fewer than 100 distinct users or 100 distinct units were grouped together prior to the initial data aggregations for files that include service code information. This grouped $213 million allowed (9% of total) under the "AGGRGTDSVC = Aggregated Services" code before file-specific generalization algorithms are applied.
- A provider must have paid claims for at least 25 distinct members for a given payer for provider-specific data to be reported for that provider under that payer code. This

condition results in $444.8 million allowed (19% of total) being grouped *a priori* as "Small Providers."

- To reduce the amount of generalization required, the Chronic Condition Hierarchical Grouping field (CCHG_Cat, a diagnosis-driven categorization of members calculated by the NH CHIS data vendor) was simplified as follows by creating single "Healthy Male" and "Healthy Female" categories for members 65 and over:

Old Value	Old Value Description	New Value	New Value Description
131	Healthy Male (65-69)	145	Healthy Male (65+)
132	Healthy Male (70-74)	145	Healthy Male (65+)
133	Healthy Male (75-79)	145	Healthy Male (65+)
134	Healthy Male (80-84)	145	Healthy Male (65+)
135	Healthy Male (85+)	145	Healthy Male (65+)
139	Healthy Female (65-69)	146	Healthy Female (65+)
140	Healthy Female (70-74)	146	Healthy Female (65+)
141	Healthy Female (75-79)	146	Healthy Female (65+)
142	Healthy Female (80-84)	146	Healthy Female (65+)
143	Healthy Female (85+)	146	Healthy Female (65+)

- Additionally, null Chronic Condition Hierarchical Grouping values and the Chronic Condition Hierarchical Grouping value "144 = Other Healthy (Unknown Age or Gender)" were grouped into the BerryDunn-created Chronic Condition Hierarchical Grouping value "999 = "Null, Unknown, or Masked CCHG." No other Chronic Condition Hierarchical Grouping values were recoded prior to equivalence class testing. Also, for members with claims associated with multiple Chronic Condition Hierarchical Groupings, all claims for a given member in the test file were associated with a single Chronic Condition Hierarchical Grouping using the following algorithm:
    - If a member had claims associated with only one non-null Chronic Condition Hierarchical Grouping, all claims for that member were associated with that non-null Chronic Condition Hierarchical Grouping (that is, the null-value Chronic Condition Hierarchical Grouping values were recoded with the non-null Chronic Condition Hierarchical Grouping appearing for some of the member's other claims).
    - If a member had multiple valid Chronic Condition Hierarchical Grouping reported, all claims for the member were recoded with the Chronic Condition Hierarchical Grouping appearing most recently (that is, with the latest service date).
- To reduce the amount of generalization required by aggregating some rarer values, the product type variable (PRODUCT_TYPE, ME003) was simplified as follows:

Old Value	New Value	New Value Description
HMO	HMO	Health Maintenance Organization Product
PPO	PPO/POS	Preferred Provider Org. or Point-of-Service Product
POS	PPO/POS	Preferred Provider Org. or Point-of-Service Product
All other Values	OUM	Other, Unknown, or Masked

- To reduce the amount of generalization required by aggregating some rarer values, the market segment variable (MARKET_CAT, ME030) was simplified as follows:

Old Value	New Value	New Value Description
GLG1	GLG1	Large group, 51-100 members
GLG2	GLG2	Large group, 101+ members
GS1	GSG	Small group, 1-50 members
GS2	GSG	Small group, 1-50 members
GS3	GSG	Small group, 1-50 members
GS4	GSG	Small group, 1-50 members
IND	IND	Individual policy
All other Values	OUM	Other, Unknown, or Masked

- To reduce the amount of generalization required by aggregating some rarer values, the coverage type (COV_TYPE, or funding source, ME029) variable was simplified as follows:

Old Value	New Value	New Value Description
ASO	SI	Self-funded
ASW	SI	Self-funded
UND	FI	Fully-insured
All other Values	OUM	Other, Unknown, or Masked

- To reduce the amount of generalization required by aggregating some rarer values, the brand status (BRAND_STATUS, or funding source, PC029) variable was simplified as follows:

Old Value	Old Value Description	New Value	New Value Description
GENERIC	Generic	GENERIC	Generic
MSB	Multi-source Brand	BRAND	Brand
SSB	Single Source Brand	BRAND	Brand
All Other Values		OUM	Other, Unknown, or Masked

BerryDunn developed the following aggregated demographic concepts for use in the files:

- Member age (AGE, ME014, DC035, MC059, PC032) was grouped as follows for files including member demographics:

Age Group	Age Group Description
1	Ages 0-25
2	Ages 26-64
3	Ages 65+

- The following New Hampshire geographic region indicator, NH_REGION, was developed from values of MEMBER_COUNTY (MC016):

NH Region	NH Region Description
1	Seacoast, Portsmouth, & Dover: Strafford and Rockingham Counties
2	Manchester, Nashua, & Concord: Merrimack and Hillsborough Counties
3	Rural Counties: Belknap, Carroll, Cheshire, Coos, Grafton, & Sullivan
999	Outside New Hampshire

BerryDunn also added the field "UTILIZATION_TYPE" to the medical claims. This value-added service categorization is derived from the place of service (POS, MC037) field on professional claims and the type of bill (UB_BILL_TYPE, MC036) on facility claims, and takes on the values:

- Ambulatory Surgery
- Clinic/Office
- Home Health/Hospice
- Hospital Outpatient
- Hospital Inpatient
- Hospital Inpatient Professional Services
- Nursing Facility
- OUM = Other, Unknown, or Masked

BerryDunn has provided its programming code to DHHS, NHID, and the departments' data vendor to provide detailed documentation of these transformations and facilitate initial production of the enhanced public use files.

The following sections describe the specific methodologies for the nine aggregated and two person-level files comprising the enhanced public use file design.

Table 1

Medical Expenses and Total Patients by Payer, Provider, Service Category, and Member Region (File A1)

Row Definition Fields	Order of Generalization	Field Name	Common Name	NH CHIS Data Element ID	Type	Length	Description	On Current Public Use Files?
✓		service_year	Service Year	MC059, PC032	Number	4		Yes
✓	6	parent_payer_code	Parent Payer Code	MC001, PC001	String	8	This field is the Payer ID Number of the data submitter company that links to the REF_PAYER file using the Parent_Payer_Code value. This code is used to identify the data reporter. It is based upon MC001, PC001, or DC001.	No
	6	company	Company Name	N/A	String	TBD	Name of the data submitting company associated with the parent_payer_code in the NH Consolidated CHIS REF_PAYER table (e.g., Anthem Health Plans of New Hampshire, Inc.)	No
✓	7	uf_prim_elig	Primary Eligibility Indicator	ME028	String	1	Y/N field indicating whether claims summarized on the row are for members' primary health insurance eligibility ("Y") or not ("N")	Yes
✓	3	npi	National Provider Identifier	MC026	String	10	Service Provider's National Provider Identifier (NPI), assigned by the CMS National Plan and Provider Enumeration System (NPPES). A provider must have served at least 25 distinct members for a given payer to be reported individually for that payer. NPIs are masked for payer/provider combinatinos not meeting this minimum and grouped as "small payers."	No
	3	provider_name	Provider Name	MC030, MC028, MC029	String	255	Name of facility or practitioner providing the service, as reported by the carrier. A provider must have served at least 25 distinct members for a given payer to be reported individually for that payer. NPIs are masked for payer/provider combinatinos not meeting this minimum and grouped as "small payers."	Yes
	4	provicer_city	Provider Location City	MC033, PC022, DC027	String	2	City where service provider is located, as reported by the carrier.	Yes
	5	provider_state	Provider Location State	MC034, PC023, DC028	String	2	State where service provider is located, as reported by the carrier.	Yes
✓	2	MR_LINE_DESC2	Service Category	N/A	String	50	Service Category; the high-level description for the Milliman HCG code (MR_line, linked through MR_line_case_key on medical claims) associated with the claim line.	Yes
✓	1	NH_Res	Member NH Residency Indicator	MC015, PC015	String	1	Indicator variable for member NH residence. Valid values are: 0…Member is not an NH resident 1…Member is an NH resident 999…Unknown or masked by generalization algorithm	Yes
✓	1	NH_Region	NH Geographical Area	MC016, PC016	String	TBD	NH geographical region, a value-added field derived from Member County on the NH Consolidated CHIS. Valid values are: 1…Seacoast, Portsmouth, & Dover: Strafford and Rockingham Counties 2…Manchester, Nashua, & Concord: Merrimack and Hillsborough Counties 3…Rural Counties: Belknap, Carroll, Cheshire, Coos, Grafton, & Sullivan 999…Not New Hampshire, unknown, or masked by generalization algorithm	Derivable

Table 1 (cont'd)

Medical Expenses and Total Patients by Payer, Provider, Service Category, and Member Region (File A1)

Row Definition Fields	Order of Generalization	Field Name	Common Name	NH CHIS Data Element ID	Type	Length	Description	On Current Public Use Files?
		total_billed	Total billed	MC062, PC035	Decimal	10,2	This field identifies the sum of all amounts billed by providers to the plans in the reporting year. This is a money field containing dollars and cents.	Not Derivable by Payer
		total_allowed	Total Allowed	MC063, MC067, MC066, MC065, PC036, PC042, PC041, PC040	Decimal	10,2	This field identifies the sum of all plan and member payments for services rendered to the individual in the reporting year. This is a money field containing dollars and cents.	Not Derivable by Payer
		total_paid	Total Paid	MC063, PC036, MC064	Decimal	10,2	This field identifies the sum of all plan payments for services rendered in the reporting year. This is a money field containing dollars and cents.	Not Derivable by Payer
		total_deduct	Total Deductible Amounts	MC067, PC042	Decimal	10,2	This field identifies the sum of all member deductible payments for services rendered in the reporting year. This is a money field containing dollars and cents.	Not Derivable by Payer
		total_coins	Total Coinsurance Amounts	MC066, PC041	Decimal	10,2	This field identifies the sum of all member coinsurance payments (the percentage amount a member must pay for a covered service after any deductible amount is met) for services rendered in the reporting year. This is a money field containing dollars and cents.	Not Derivable by Payer
		total_copay	Total Copay Amounts	MC066, PC041	Decimal	10,2	This field identifies the sum of all member copay payments (preset, fixed dollar amounts the member is responsible for, usually on a per-visit/per-service basis, after any deductible amounts are met) for services rendered in the reporting year. This is a money field containing dollars and cents.	Not Derivable by Payer
		claim_line_count	Count of Claim Lines Summarized per Row	N/A	Number	10	Number of paid claim line items summarized in the present row. Paid claim line items (SV_STAT = "P") are summed with a value of 1. Reversals (SV_STAT = "R") are summed with a value of -1.	Not Derivable by Payer
		distinct_users	Uniique Users	N/A	Number	10	Unique members utilizing paid services summarized on a given row. No value < 11 can be reported in this field.	No
		total_patients	Total Patients	N/A	Number	10	Sum of total patients utilizing paid services summarized on a given row. Total_patients will be duplicative of unique users where a single unique user (as defined by person_key as reported in NH CHIS) has received multiple services summarized on a single generalized row. The value of total_patients should be equal to or greater than the value of distinct_users. No value < 11 can be reported in this field.	No
		generalized_row	Generalized Row Indicator	N/A	String	1	Derived field indicating whether the data in the row have been affected by the generalization algorithm ("Y"), or reflect an aggregation of the data as reported in the CHIS claims files ("N") without generalization.	No

4 Aggregated Files

The proposed design includes the following nine aggregated data files:

- Medical Expense by Payer, Provider, Service Category, and Member Region (File A1)
- Payer Expense by Member Demographics and Product Type, Medical and Pharmacy (Files A2 – A3)
- Medical, Pharmacy, and Dental Membership (Files A4 – A6)
- Medical Expense and Patients Served by Service and Primary Diagnosis (Files A7 – A8)
- Pharmacy Expense and Patients Served by Drug (File A9)

The following seven subsections describe the designs, calculation, de-identification methodology, and the adequacy of the de-identification methodology and results for each file.

4.1 Medical Expense and Members Served by Payer, Provider, Service Category, and Member Region (File A1)

Table 1 above shows the proposed layout of the medical expense by payer, provider, service category, and member region summary file (File A1). As aggregated data, the file does not comprise personal health information. However, to ensure individual privacy, the methodology follows the CMS cell suppression standards, which prohibit any value greater than 0 and less than 11 to be displayed for any measure of services or patients (any amount is considered reportable in dollar fields), by applying generalization or suppression where necessary to avoid the display of too-small claim line or patient counts.

The file initially aggregates all medical claim dollars, claim lines, and patients served by the following "group-by" fields, as they appear in the base test data:

- Payer, as defined by the NH CHIS data field PARENT_PAYER_CODE. As noted above in the general methodology section, any payer code that did not account for at least 1% of paid claim dollar volume was aggregated into a new payer code, "OTHPAYR = Other or Unknown Payer" before the initial aggregation. For medical, this also happened to be the universe of payers with 50,000 member months or more. The payer company name (COMPANY) is joined to the claim results from the Ref_Payer table.
- Primary Eligibility, as defined by the NH CHIS data field UF_PRIM_ELIG ("Primary Eligibility Use Flag"). The field is coded "Y" if the claims summarized on a given row are associated with the members' primary health insurance eligibility, "N" otherwise.
- NPI, National Provider Identifier. NPIs (NPI, MC026, MP017) are joined to claims from the Provider_Detail table prior to the initial aggregation. If a valid SERV_PROV_KEY (NH CHIS warehouse-level provider identifier) is reported on the claim, SERV_PROV_KEY on the claim table is joined to PROV_KEY (MC024, MC076, MP003) on Provider_Detail. If a claim line

reports a valid BILL_PROV_KEY but no valid SERV_PROV_KEY, the join is made using BILL_PROV_KEY and PROV_KEY.

- Provider Name is joined to the claim file before the initial aggregation as described above for NPI.
- Provider City, PROV_CLINIC_CITY (MC033), is joined to the claim file before the initial aggregation as described above for NPI.
- Provider State, PROV_CLINIC_STATE (MC034), is joined to the claim file before the initial aggregation as described above for NPI.
- MR_Line_Desc2, the service class description MR_LINE_DESC_2 from the Ref_HCG table, is joined to the claims before the initial aggregation by values of the Health Cost Grouper[xiii] (HCG) field MR_LINE_CASE_KEY appearing on both the claims and Ref_HCG. MR_LINE_DESC_2 provides information on the type of service rendered for lines where the service code must be generalized to meet cell-reporting guidelines. Please note that MR_LINE_DESC_2 and SERVICE_CODE have a many-to-many relationship.[xiv]
- NH Resident: Based on values of MEM_STATE (ME016), the NH_RES field takes on the value "1" if the reported member state is New Hampshire, "0" otherwise.
- NH Region, as described in the general methodology section.

Generalized Row, a value-added field created after the generalization process, is included on the final output file. A "Y" in this field indicates that some or all of the records summarized on the row required at least some generalization to be reportable. An "N" indicates all claim lines summarized on the row were reportable under the values appearing for the constituent rows in the test data.

In testing, claims representing $565 million (24%) of allowed charges volume summed to unreportable rows (rows with less than 11 claim lines or users) after the initial aggregation. BerryDunn applied the following generalization algorithm to address the unreportable cells. Because this file is the only file that includes provider data elements, provider information was treated as the most important data to maintain where possible. After each of the numbered generalizations, the data were re-summarized by the "group-by" variables and the rows re-checked for adequate cell size.

[xiii] Health Cost Groupers are a value-added set of claim groupers developed by the NH CHIS data vendor.

[xiv] However, this does not result in double counting because each claim line is associated with exactly one SERVICE_CODE and MR_LINE_DESC_2, and the join of MR_LINE_CASE_KEY between claims and Ref_HCG is many-to-one.

Generalization order:

1. Geography (NH residence and region): NH_Region was recoded as "999" and NH_State as "0" for all rows failing the cell suppression standards before generalization. This generalization rendered 21% ($120 million allowed) of the unreportable volume reportable, leaving $445 million (19% of commercial medical claims total) that required further generalization.
2. Service Category: MR_LINE_DESC_2 was recoded as "Generalized" for rows failing the cell suppression standards, and the aggregation was repeated. This generalization rendered 82% ($366 million allowed) of the unreportable volume reportable, leaving $78 million allowed (3% of commercial medical claims total) that required further generalization.
3. Provider NPI & Name: NPI and provider name were recoded as "9999999999" and "Small Provider," respectively, for the remaining rows failing the cell suppression standards, and the aggregation was repeated. This generalization rendered 24% ($19 million allowed) of the unreportable volume reportable, leaving $60 million allowed (2.5% of commercial medical claims total) that required further generalization.
4. Provider City: provider city was recoded as "Generalized" for the remaining rows failing the cell suppression standards, and the aggregation was repeated. This generalization rendered 97% ($58 million allowed) of the unreportable volume reportable, leaving $1.8 million allowed (0.1% of commercial medical claims total) that required further generalization.
5. Provider State: provider state was recoded as "99" for the remaining rows failing the cell suppression standards, and the aggregation was repeated. This generalization rendered 99.96% ($1.8 million allowed) of the unreportable volume reportable, leaving an immaterial $675 allowed that required further generalization.
6. Payers: PARENT_PAYER_CODE was recoded as "OTHPAYR = Other Payers" for the remaining rows with inadequate claim line and patient totals. This generalization did not result in adequate cell size for the unreportable volume.
7. Primary Eligibility: UF_PRIM_ELIG was recoded as "U = Unknown" for the remaining rows with inadequate claim line and patient totals. This generalization did not result in adequate cell size for the unreportable volume.

The remaining rows with users or claim line counts less than 11 (but not equal to 0) remaining at this point would need to be suppressed (excluded) from the published files. A summary of dollars for suppressed volume should be provided in the companion documentation if suppression is required in production.

Similar summaries of suppressed claim line counts and users should *not* be reported. If the totals for these fields do not reach the reportability threshold, such a summary would fail the reportability standards; if, on the other hand, the totals are reported where they do meet the

threshold, their absence in the companion documentation to other files (where the threshold was not met) would be tantamount to reporting totals beneath the threshold.

The final test file includes approximately 82,000 rows.

Two distinct concepts of service user counts are included in the aggregated files. DISTINCT_USERS provides a count of distinct values of PERSON_KEY represented on all claim lines summarized on the row, regardless of the level of generalization.

The TOTAL_PATIENTS field, however, counts all users for all payer/provider/patient/service category/geography combinations summarized on the row. That is, if the same member receives services in two different service categories reported on rows requiring service category generalization, those rows will add 1 to the DISTINCT_USERS count, and 2 to the TOTAL_PATIENTS count.

TOTAL_PATIENTS is calculated by summing the value of DISTINCT_USERS calculated in the initial, pre-generalization aggregation. Therefore, DISTINCT_USERS and TOTAL_PATIENTS are equal for ungeneralized rows (GENERALIZED_ROW = “N”), but TOTAL_PATIENTS will be greater than or equal to DISTINCT_USERS for generalized rows (GENERALIZED_ROW = “Y”).

BerryDunn proposes this definition for the TOTAL_PATIENTS field as a useful adjunct to DISTINCT_USERS that provides a more accurate reflection of total service utilization in the commercial market than is reflected in the distinct users count for highly generalized rows.

BerryDunn has provided its programming code to DHHS, NHID, and the departments’ data vendor to provide detailed documentation of this methodology and facilitate initial production of the enhanced public use files.

Table 2
Medical Cost and Members Served by Payer and Product Type (File A2)

Row Definition Fields	Order of Generalization	Field Name	Common Name	NH CHIS Data Element ID	Type	Length	Description	On Current Public Use Files?
✓		service_year	Service Year	MC059, PC032	Number	4		Yes
		coverage_class	Coverage Class	MC899, PC899	String	3	This field indicated the type of record. For all medical claims records, this value will be MED. Pharmacy Claims are PHM. This field is not part of the row definition, as medical, dental, and pharmacy reports are presented separately.	Yes
✓	8	parent_payer_code	Parent Payer Code	MC001, PC001	String	8	This field is the Payer ID Number of the data submitter company that links to the REF_PAYER file using the Parent_Payer_Code value. This code is used to identify the data reporter. It is based upon MC001, PC001, or DC001.	No
		company	Company Name	N/A	String	TBD	Name of the data submitting company associated with the parent_payer_code in the NH Consolidated CHIS REF_PAYER table (e.g., Anthem Health Plans of New Hampshire, Inc.)	No
✓	9	uf_prim_elig	Primary Eligibility Indicator	ME028	String	1	Y/N field indicating whether claims summarized on the row are for members' primary health insurance eligibility ("Y") or not ("N")	Yes
✓	6	FI_SI	Fully Insured/Self Insured Indicator	ME029	String	3	This field indicates the type of coverage and is used to distinguish self-funded plans from commercially insured plans as reported in ME029. FI_SI applies a simplifying recoding schema to ME029 **(see Recodes section in the documentation for the mapping)**. Valid values of FI_SI are: FI...commercially insured plan SI...self-funded plan OUM..."Other, Unknown, or Masked:" other plan type, unknown plan type, or value masked by generalization algorithm	No
✓	4	Prod_Type	Insurance Product Type	N/A			This field identifies standardized product types, and reflects a simplifying recoding schema applied to the "Product_Type" field in the NH Consolidated CHIS (**see Recodes section in the documentation for the mapping**). Valid values are: PPO/POS...Commercial Preferred Provider Organization or Point-of-Service product HMO...Commercial HMO OUM..."Other, Unknown, or Masked:" other product type, unknown product type, or value masked by generalization algorithm	Yes
✓	5	Mkt_Seg	Market Segment Code	ME030	String	4	This field indicates the size of the group for group policies, or individual market for individual policies, and applies a simplifying schema to ME030, Market Category Code (**see Recodes section in the documentation for the mapping**). Valid values of Mkt_Seg are: IND...For policies sold and issued directly to individuals, other than those sold on a franchise basis, as defined pursuant to RSA 415:19, or a group conversion policies required pursuant to RSA 415:18 VII (a) GSG...For policies sold and issued directly to employers having between1 and 50 employees GLG1...For policies sold and issued directly to employers having between 51 and 99 employees GLG2...For policies sold and issued directly to employers having 100 or more employees OUM..."Other, Unknown, or Masked:" other market segment, unknown market segment, or value masked by generalization algorithm	No

Table 2 (cont'd)
Medical Cost and Members Served by Payer and Product Type (File A2)

Row Definition Fields	Order of Generalization	Field Name	Common Name	NH CHIS Data Element ID	Type	Length	Description	On Current Public Use Files?
✓	7	Utilization_Category	Utilization Category	MC036, MC037	String	60	This field identifies the utilization category for this record. It is a value-added field based on the Type of Bill field for institutional claims and Place of Service for professional claims. Valid values are: Ambulatory Surgery Clinic/Office Home Health/Hospice Hospital Inpatient Hospital Inpatient Professional Services Hospital Outpatient Nursing Facility OUM	Derivable
✓	3	gender	Member Gender	MC012	String	1	M/F	Yes
✓	1	age_group_code	Member Age Group Code	ME014, MC059, PC032	Number	1	Numeric code associated with age groupings derived from member age fields in the NH Consolidated CHIS. Valied values are: 1…Ages 0-25 2…Ages 26-64 3…Ages 65+ 999…Unknown or masked by generalization algorithm	Derivable
✓	2	NH_Res	Member NH Residency Indicator	MC015, PC015	String	1	Indicator variable for member NH residence. Valid values are: 0…Member is not an NH resident 1…Member is an NH resident 999…Unknown or masked by generalization algorithm	Yes
✓	2	NH_Region	NH Geographical Area	MC016, PC016	String	TBD	NH geographical region, a value-added field derived from Member County on the NH Consolidated CHIS. Valid values are: 1…Seacoast, Portsmouth, & Dover: Strafford and Rockingham Counties 2…Manchester, Nashua, & Concord: Merrimack and Hillsborough Counties 3…Rural Counties: Belknap, Carroll, Cheshire, Coos, Grafton, & Sullivan 999…Not New Hampshire, unknown, or masked by generalization algorithm	Derivable

Table 2 (cont'd)
Medical Cost and Members Served by Payer and Product Type (File A2)

Row Definition Fields	Order of Generalization	Field Name	Common Name	NH CHIS Data Element ID	Type	Length	Description	On Current Public Use Files?
		total_billed	Total billed	MC062, PC035	Decimal	10,2	This field identifies the sum of all amounts billed by providers to the plans in the reporting year. This is a money field containing dollars and cents.	Not Derivable by Payer
		total_allowed	Total Allowed	MC063, MC067, MC066, MC065, PC036, PC042, PC041, PC040	Decimal	10,2	This field identifies the sum of all plan and member payments for services rendered to the individual in the reporting year. This is a money field containing dollars and cents.	Not Derivable by Payer
		total_paid	Total Paid	MC063, PC036, MC064	Decimal	10,2	This field identifies the sum of all plan payments for services rendered in the reporting year. This is a money field containing dollars and cents.	Not Derivable by Payer
		total_deduct	Total Deductible Amounts	MC067, PC042	Decimal	10,2	This field identifies the sum of all member deductible payments for services rendered in the reporting year. This is a money field containing dollars and cents.	Not Derivable by Payer
		total_coins	Total Coinsurance Amounts	MC066, PC041	Decimal	10,2	This field identifies the sum of all member coinsurance payments (the percentage amount a member must pay for a covered service after any deductible amount is met) for services rendered in the reporting year. This is a money field containing dollars and cents.	Not Derivable by Payer
		total_copay	Total Copay Amounts	MC066, PC041	Decimal	10,2	This field identifies the sum of all member copay payments (preset, fixed dollar amounts the member is responsible for, usually on a per-visit/per-service basis, after any deductible amounts are met) for services rendered in the reporting year. This is a money field containing dollars and cents.	Not Derivable by Payer
		claim_line_count	Count of Claim Lines Summarized per Row	N/A	Number	10	Number of paid claim line items summarized in the present row. Paid claim line items (SV_STAT = "P") are summed with a value of 1. Reversals (SV_STAT = "R") are summed with a value of -1.	Not Derivable by Payer
		distinct_users	Uniique Users	N/A	Number	10	Unique members utilizing paid services summarized on a given row. No value < 11 can be reported in this field.	No
		total_patients	Total Patients	N/A	Number	10	Sum of total patients utilizing paid services summarized on a given row. Total_patients will be duplicative of unique users where a single unique user (as defined by person_key as reported in NH CHIS) has received multiple services summarized on a single generalized row. The value of total_patients should be equal to or greater than the value of distinct_users. No value < 11 can be reported in this field.	No
		generalized_row	Generalized Row Indicator	N/A	String	1	Derived field indicating whether the data in the row have been affected by the generalization algorithm ("Y"), or reflect an aggregation of the data as reported in the CHIS claims files ("N") without generalization.	No

4.2 Medical Expense and Members Served by Product Type, Market Segment, and Funding Type (File A2)

Table 2 above shows the proposed layout of the medical expense by product type, market segment, funding type, and member demographics summary file (File A2). As aggregated data, the file does not comprise personal health information. However, to ensure individual privacy, the methodology follows the CMS cell suppression standards, which prohibit any value greater than 0 and less than 11 to be displayed for any measure of services or patients (any amount is considered reportable in dollar fields), by applying generalization or suppression where necessary to avoid the display of too-small claim or patient counts.

The file initially aggregates all medical claim dollars, claim lines, and patients served by the following "group-by" fields, as they appear in the base test data:

- Payer, as defined by PARENT_PAYER_CODE. As noted above in the general methodology section, any payer code that did not account for at least 1% of paid claim dollar volume was aggregated into a new payer code, "OTHPAYR = Other or Unknown Payer" before the initial aggregation. For medical, this also happened to be the universe of payers with 50,000 member months or more. The payer company name (COMPANY) is joined to the claim results from the Ref_Payer table.
- Primary Eligibility, as defined by the NH CHIS data field UF_PRIM_ELIG ("Primary Eligibility Use Flag"). The field is coded "Y" if the claims summarized on a given row are associated with the members' primary health insurance eligibility, "N" otherwise.
- FI/SI: The FI_SI field is the result of the simplification algorithm described in the general methodology section for the Consolidated NH CHIS field COV_TYPE. This information is joined onto the claims from the Membership_MC_2016 table by the field MEMBER_MONTH_KEY, which is present on both the claims and membership tables.
- Product Type: PROD_TYPE is the result of the simplification algorithm described in the general methodology section for the Consolidated NH CHIS field PRODUCT_TYPE.
- Market Segment: MKT_SEG is the result of the simplification algorithm described in the general methodology section for the Consolidated NH CHIS field COV_TYPE. This information is joined onto the claims from the Membership_MC_2016 table by the field MEMBER_MONTH_KEY, which is present on both the claims and membership tables.
- Utilization Type: UTILIZATION_TYPE is the eight-category utilization type indicator developed by BerryDunn, as described in the general methodology section.
- Gender: Member gender, as reported in the NH CHIS field SEX (MC012). Valid values are "F," "M," and "U = Unknown or Masked."
- Age group, as described in the general methodology section.

- NH Resident: Based on values of MEM_STATE (ME016), the NH_RES field takes on the value "1" if the reported member state is New Hampshire, "0" otherwise.
- NH Region, as described in the general methodology section.

Generalized Row, a value-added field created after the generalization process, is included on the final output file. A "Y" in this field indicates that some or all of the records summarized on the row required at least some generalization to be reportable. An "N" indicates all claim lines summarized on the row were reportable under the values appearing for the constituent rows in the test data.

In testing, claims representing $75 million (3%) of allowed charges volume summed to unreportable rows (rows with less than 11 units or users) after the initial aggregation. BerryDunn applied the following generalization algorithm to address the unreportable cells. After each of the numbered generalizations, the data were re-summarized by the "group-by" variables and the rows re-checked for adequate cell size:

1. Age group: AGE_GROUP was recoded as "999 = Unknown or Masked" for all rows failing the cell reporting standards after the initial aggregation. This generalization rendered 24% ($18 million allowed) of the unreportable volume reportable, leaving $57 million allowed (2% of commercial medical claims total) that required further generalization.
2. Geography (NH residence and region): NH_Region was recoded as "999" and NH_State as "0" for all rows failing the cell suppression standards after age generalization. This generalization rendered 69% ($39 million allowed) of the remaining unreportable volume reportable, leaving $18 million (1% of commercial medical claims total) that required further generalization.
3. Gender: Gender was recoded as "U" for the remaining rows with inadequate claim line and patient totals. This generalization resolved 37% ($6.5 million allowed) of the remaining unreportable volume reportable, leaving $11 million (0.5% of commercial medical claims total) that required further generalization.
4. Product Type: PROD_TYPE was recoded as "OUM = Other, Unknown, or Masked" for the remaining rows with inadequate claim line and patients totals. This generalization resolved 21% ($2.3 million allowed) of the remaining unreportable volume, leaving $8.8 million (0.4% of commercial medical claims total) that required further generalization.
5. Market Segment: MKT_SEG was recoded as "OUM = Other, Unknown, or Masked" for the remaining rows with inadequate claim line and patient totals. This generalization resolved 50% ($4.4 million allowed) of the remaining unreportable volume, leaving $4.4 million (0.2% of commercial medical claims total) that required further generalization.
6. Funding Type: FI_SI was recoded as "OUM = Other, Unknown, or Masked" for the remaining rows with inadequate claim line and patient totals. This generalization resolved 64% ($2.8

million allowed) of the remaining unreportable volume, leaving $1.6 million (0.1% of commercial medical claims total) that required further generalization.

7. Utilization Type: UTILIZATION_TYPE was recoded as "OUM = Other, Unknown, or Masked" for the remaining rows with inadequate claim line and patient totals. This generalization resolved 99% ($1.58 million allowed) of the remaining unreportable volume, leaving an immaterial $14,613 that required further generalization.
8. Payers: PARENT_PAYER_CODE was recoded as "OTHPAYR = Other Payers" for the remaining rows with inadequate units and patients totals. This generalization rendered all claim volume reportable.
9. Primary Eligibility: Had any unreportable volume remained, UF_PRIM_ELIG would have been recoded as "U = Unknown" for the remaining rows with inadequate claim line and patient totals.

Although this generalization algorithm results in no suppressed volume in testing, *any rows with user or claim counts less than 11 (but not equal to 0) remaining at this point would need to be suppressed (excluded) from the published files.*

A summary of dollars for suppressed volume should be provided in the companion documentation if suppression is required in production.

Similar summaries of suppressed claim line counts and users should *not* be reported. If the totals for these fields do not reach the reportability threshold, such a summary would fail the reportability standards; if, on the other hand, the totals are reported where they do meet the threshold, their absence in the companion documentation to other files (where the threshold was not met) would be tantamount to reporting totals beneath the threshold.

The final test file includes approximately 6,600 rows.

Two distinct concepts of service user counts are included in the aggregated files. DISTINCT_USERS provides a count of distinct values of PERSON_KEY represented on all claim lines summarized on the row, regardless of the level of generalization.

The TOTAL_PATIENTS field, however, counts all users for all payer/product type/utilization type/member demographic combinations summarized on the row. That is, if the same member receives services in two different utilization types reported on rows requiring utilization type generalization, those rows will add 1 to the DISTINCT_USERS count, and 2 to the TOTAL_PATIENTS count.

TOTAL_PATIENTS is calculated by summing the value of DISTINCT_USERS calculated in the initial, pre-generalization aggregation. Therefore, DISTINCT_USERS and TOTAL_PATIENTS are equal for ungeneralized rows (GENERALIZED_ROW = "N"), but TOTAL_PATIENTS will be greater than or equal to DISTINCT_USERS for generalized rows (GENERALIZED_ROW = "Y").

BerryDunn proposes this definition for the TOTAL_PATIENTS field as a useful adjunct to DISTINCT_USERS that provides a more accurate reflection of total service utilization in the commercial market than is reflected in the distinct users count for highly generalized rows.

BerryDunn has provided its programming code to DHHS, NHID, and the departments' data vendor to provide detailed documentation of this methodology and facilitate initial production of the enhanced public use files.

Table 3
Pharmacy Cost and Members Served by Payer and Product Type (File A3)

Row Definition Fields	Order of Generalization	Field Name	Common Name	NH CHIS Data Element ID	Type	Length	Description	On Current Public Use Files?
✓		service_year	Service Year	MC059, PC032	Number	4		Yes
		coverage_class	Coverage Class	MC899, PC899	String	3	This field indicated the type of record. For all medical claims records, this value will be MED. Pharmacy Claims are PHM. This field is not part of the row definition, as medical, dental, and pharmacy reports are presented separately.	Yes
✓	8	parent_payer_code	Parent Payer Code	MC001, PC001	String	8	This field is the Payer ID Number of the data submitter company that links to the REF_PAYER file using the Parent_Payer_Code value. This code is used to identify the data reporter. It is based upon MC001, PC001, or DC001.	No
		company	Company Name	N/A	String	TBD	Name of the data submitting company associated with the parent_payer_code in the NH Consolidated CHIS REF_PAYER table (e.g., Anthem Health Plans of New Hampshire, Inc.)	No
✓	9	uf_prim_elig	Primary Eligibility Indicator	ME028	String	1	Y/N field indicating whether claims summarized on the row are for members' primary health insurance eligibility ("Y") or not ("N")	Yes
✓	6	FI_SI	Fully Insured/Self Insured Indicator	ME029	String	3	This field indicates the type of coverage and is used to distinguish self-funded plans from commercially insured plans as reported in ME029. FI_SI applies a simplifying recoding schema to ME029 **(see Recodes section in the documentation for the mapping)**. Valid values of FI_SI are: FI…commercially insured plan SI…self-funded plan OUM…"Other, Unknown, or Masked:" other plan type, unknown plan type, or value masked by generalization algorithm	No
✓	4	Prod_Type	Insurance Product Type	N/A			This field identifies standardized product types, and reflects a simplifying recoding schema applied to the "Product_Type" field in the NH Consolidated CHIS (**see Recodes section in the documentation for the mapping**). Valid values are: PPO/POS…Commercial Preferred Provider Organization or Point-of-Service product HMO…Commercial HMO OUM…"Other, Unknown, or Masked:" other product type, unknown product type, or value masked by generalization algorithm	Yes
✓	5	Mkt_Seg	Market Segment Code	ME030	String	4	This field indicates the size of the group for group policies, or individual market for individual policies, and applies a simplifying schema to ME030, Market Category Code (**see Recodes section in the documentation for the mapping**). Valid values of Mkt_Seg are: IND…For policies sold and issued directly to individuals, other than those sold on a franchise basis, as defined pursuant to RSA 415:19, or a group conversion policies required pursuant to RSA 415:18 VII (a) GSG…For policies sold and issued directly to employers having between1 and 50 employees GLG1…For policies sold and issued directly to employers having between 51 and 99 employees GLG2…For policies sold and issued directly to employers having 100 or more employees OUM…"Other, Unknown, or Masked:" other market segment, unknown market segment, or value masked by generalization algorithm	No

Table 3 (cont'd)
Pharmacy Cost and Members Served by Payer and Product Type (File A3)

Row Definition Fields	Order of Generalization	Field Name	Common Name	NH CHIS Data Element ID	Type	Length	Description	On Current Public Use Files?
✓	7	Drug_Ther_Class	Drug Therapeutic Class	N/A	String	58	High-level NDC therapeutic class description (field name "Main"), joined to the pharmacy claims by the field NDC_THER_CLASS (calculated from MC075, PC026) from the Consolidated CHIS REF_NDC_THERAPEUTIC_CLASS reference table.	No
✓	3	gender	Member Gender	MC012	String	1	M/F	Yes
✓	1	age_group_code	Member Age Group Code	ME014, MC059, PC032	Number	1	Numeric code associated with age groupings derived from member age fields in the NH Consolidated CHIS. Valied values are: 1…Ages 0-25 2…Ages 26-64 3…Ages 65+ 999…Unknown or masked by generalization algorithm	Derivable
✓	2	NH_Res	Member NH Residency Indicator	MC015, PC015	String	1	Indicator variable for member NH residence. Valid values are: 0…Member is not an NH resident 1…Member is an NH resident 999…Unknown or masked by generalization algorithm	Yes
✓	2	NH_Region	NH Geographical Area	MC016, PC016	String	TBD	NH geographical region, a value-added field derived from Member County on the NH Consolidated CHIS. Valid values are: 1…Seacoast, Portsmouth, & Dover: Strafford and Rockingham Counties 2…Manchester, Nashua, & Concord: Merrimack and Hillsborough Counties 3…Rural Counties: Belknap, Carroll, Cheshire, Coos, Grafton, & Sullivan 999…Not New Hampshire, unknown, or masked by generalization algorithm	Derivable
		total_billed	Total billed	MC062, PC035	Decimal	10,2	This field identifies the sum of all amounts billed by providers to the plans in the reporting year. This is a money field containing dollars and cents.	Not Derivable by Payer
		total_allowed	Total Allowed	MC063, MC067, MC066, MC065, PC036, PC042, PC041, PC040	Decimal	10,2	This field identifies the sum of all plan and member payments for services rendered to the individual in the reporting year. This is a money field containing dollars and cents.	Not Derivable by Payer
		total_paid	Total Paid	MC063, PC036, MC064	Decimal	10,2	This field identifies the sum of all plan payments for services rendered in the reporting year. This is a money field containing dollars and cents.	Not Derivable by Payer
		total_deduct	Total Deductible Amounts	MC067, PC042	Decimal	10,2	This field identifies the sum of all member deductible payments for services rendered in the reporting year. This is a money field containing dollars and cents.	Not Derivable by Payer
		total_coins	Total Coinsurance Amounts	MC066, PC041	Decimal	10,2	This field identifies the sum of all member coinsurance payments (the percentage amount a member must pay for a covered service after any deductible amount is met) for services rendered in the reporting year. This is a money field containing dollars and cents.	Not Derivable by Payer
		total_copay	Total Copay Amounts	MC066, PC041	Decimal	10,2	This field identifies the sum of all member copay payments (preset, fixed dollar amounts the member is responsible for, usually on a per-visit/per-service basis, after any deductible amounts are met) for services rendered in the reporting year. This is a money field containing dollars and cents.	Not Derivable by Payer

Table 3 (cont'd)
Pharmacy Cost and Members Served by Payer and Product Type (File A3)

Row Definition Fields	Order of Generalization	Field Name	Common Name	NH CHIS Data Element ID	Type	Length	Description	On Current Public Use Files?
		total_days_supply	Total days' supply	PC034	Decimal	10,2	This field identifies the total days' supply of paid utilization.	Not Derivable by Payer
		claim_line_count	Count of Claim Lines Summarized per Row	N/A	Number	10	Number of paid claim line items summarized in the present row. Paid claim line items (SV_STAT = "P") are summed with a value of 1. Reversals (SV_STAT = "R") are summed with a value of -1.	Not Derivable by Payer
		distinct_users	Uniique Users	N/A	Number	10	Unique members utilizing paid services summarized on a given row. No value < 11 can be reported in this field.	No
		total_patients	Total Patients	N/A	Number	10	Sum of total patients utilizing paid services summarized on a given row. Total_patients will be duplicative of unique users where a single unique user (as defined by person_key as reported in NH CHIS) has received multiple services summarized on a single generalized row. The value of total_patients should be equal to or greater than the value of distinct_users. No value < 11 can be reported in this field.	No
		generalized_row	Generalized Row Indicator	N/A	String	1	Derived field indicating whether the data in the row have been affected by the generalization algorithm ("Y"), or reflect an aggregation of the data as reported in the CHIS claims files ("N") without generalization.	No

4.3 Pharmacy Expense and Members Served by Product Type, Market Segment, and Funding Type (File A3)

Table 3 above shows the proposed layout of the pharmacy expense by product type, market segment, funding type, and member demographics summary file (File A3). As aggregated data, the file does not comprise personal health information. However, to ensure individual privacy, the methodology follows the CMS cell suppression standards, which prohibit any value greater than 0 and less than 11 to be displayed for any measure of services or patients (any amount is considered reportable in dollar fields), by applying generalization or suppression where necessary to avoid the display of too-small claim or patient counts.

In cases where pharmacy benefit manager (PBM) payers and a member's main carrier have both reported a claim, the pharmacy claim files include the data from both payers.

The file initially aggregates all pharmacy claim dollars, claim lines, drug days supply, and patients served by the following "group-by" fields, as they appear in the base test data:

- Payer, as defined by PARENT_PAYER_CODE. As noted above in the general methodology section, any payer code that did not account for at least 1% of paid claim dollar volume was aggregated into a new payer code, "OTHPAYR = Other or Unknown Payer" before the initial aggregation. The payer company name (COMPANY) is joined to the claim results from the Ref_Payer table.
- Primary Eligibility, as defined by the NH CHIS data field UF_PRIM_ELIG ("Primary Eligibility Use Flag"). The field is coded "Y" if the claims summarized on a given row are associated with the members' primary health insurance eligibility, "N" otherwise.
- FI/SI: The FI_SI field is the result of the simplification algorithm described in the general methodology section for the Consolidated NH CHIS field COV_TYPE. This information is joined onto the claims from the Membership_PC_2016 table by the field MEMBER_MONTH_KEY, which is present on both the claims and membership tables.
- Product Type: PROD_TYPE is the result of the simplification algorithm described in the general methodology section for the Consolidated NH CHIS field PRODUCT_TYPE.
- Market Segment: MKT_SEG is the result of the simplification algorithm described in the general methodology section for the Consolidated NH CHIS field COV_TYPE. This information is joined onto the claims from the Membership_PC_2016 table by the field MEMBER_MONTH_KEY, which is present on both the claims and membership tables.
- Drug Therapeutic Class: DRUG_THER_CLASS shows the values of the MAIN field in the Ref_NDC_Ther_Class table, merged onto the pharmacy claims by joining the pharmacy claims field NDC_THER_CLASS to THER_CLASS on Ref_NDC_Ther_Class.

- Gender: Member gender, as reported in the NH CHIS field SEX (PC012). Valid values are "F," "M," and "U = Unknown or Masked."
- Age group, as described in the general methodology section.
- NH Resident: Based on values of MEM_STATE (ME016), the NH_RES field takes on the value "1" if the reported member state is New Hampshire, "0" otherwise.
- NH Region, as described in the general methodology section.

Generalized Row, a value-added field created after the generalization process, is included on the final output file. A "Y" in this field indicates that some or all of the records summarized on the row required at least some generalization to be reportable. An "N" indicates all claim lines summarized on the row were reportable under the values appearing for the constituent rows in the test data.

In testing, claims representing $92 million (9%) of allowed charges volume summed to unreportable rows (rows with less than 11 claim lines, days supply, or users) after the initial aggregation. BerryDunn applied the following generalization algorithm to address the unreportable cells. After each of the numbered generalizations, the data were re-summarized by the "group-by" variables and the rows re-checked for adequate cell size:

1. Age group: AGE_GROUP was recoded as "999 = Unknown or Masked" for all rows failing the cell reporting standards after the initial aggregation. This generalization rendered 16% ($15 million allowed) of the unreportable volume reportable, leaving $77 million allowed (7.5% of commercial pharmacy claims total) that required further generalization.
2. Geography (NH residence and region): NH_Region was recoded as "999" and NH_State as "0" for all rows failing the cell suppression standards after age generalization. This generalization rendered 66% ($50 million allowed) of the remaining unreportable volume reportable, leaving $26 million (2.6% of commercial pharmacy claims total) that required further generalization.
3. Gender: Gender was recoded as "U" for the remaining rows with inadequate claim, days supply, and patient totals. This generalization resolved 23% ($6 million allowed) of the remaining unreportable volume, leaving $20 million (2% of commercial pharmacy claims allowed total) that required further generalization.
4. Product Type: PROD_TYPE was recoded as "OUM = Other, Unknown, or Masked" for the remaining rows with inadequate claim, days supply, and patient totals. This generalization resolved 21% ($4 million allowed) of the remaining unreportable volume, leaving $16 million (1.6% of commercial pharmacy claims total) that required further generalization.
5. Market Segment: MKT_SEG was recoded as "OUM = Other, Unknown, or Masked" for the remaining rows with inadequate claim, days supply, and patient totals. This generalization resolved 46% ($7.4 million allowed) of the remaining unreportable volume, leaving $8.8 million (0.9% of commercial pharmacy claims total) that required further generalization.

6. Funding Type: FI_SI was recoded as "OUM = Other, Unknown, or Masked" for the remaining rows with inadequate claim, days supply, and patient totals. This generalization resolved 45% ($4 million allowed) of the remaining unreportable volume, leaving $4.8 (0.5% of commercial pharmacy claims total) that required further generalization.
7. Drug Therapeutic Class: DRUG_THER_CLASS was recoded as "UNKNOWN OR MASKED" for the remaining rows with inadequate claim, days supply, and patient totals. This generalization resolved all but an immaterial $209 of unreportable volume.
8. Payers: PARENT_PAYER_CODE was recoded as "OTHPAYR = Other Payers" for the remaining rows with inadequate claim, days supply, and patient totals. This generalization did not resolve any unreportable rows.
9. Primary Eligibility: UF_PRIM_ELIG was recoded as "U = Unknown" for the remaining rows with inadequate claim, days supply, and patient totals. This generalization did not resolve any unreportable rows.

The remaining rows with users, claim lines, or days supply less 11 (but not equal to 0) remaining at this point would need to be suppressed (excluded) from the published files.

A summary of dollars for suppressed volume should be provided in the companion documentation if suppression is required in production.

Similar summaries of suppressed claim line counts and users should *not* be reported. If the totals for these fields do not reach the reportability threshold, such a summary would fail the reportability standards; if, on the other hand, the totals are reported where they do meet the threshold, their absence in the companion documentation to other files (where the threshold was not met) would be tantamount to reporting totals beneath the threshold.

The final test file includes approximately 27,000 rows.

Two distinct concepts of service user counts are included in the aggregated files. DISTINCT_USERS provides a count of distinct values of PERSON_KEY represented on all claim lines summarized on the row, regardless of the level of generalization.

The TOTAL_PATIENTS field, however, counts all users for all payer/product type/drug therapeutic class/member demographic combinations summarized on the row. That is, if the same member receives services in two different therapeutic classes reported on rows requiring therapeutic class generalization, those rows will add 1 to the DISTINCT_USERS count, and 2 to the TOTAL_PATIENTS count.

TOTAL_PATIENTS is calculated by summing the value of DISTINCT_USERS calculated in the initial, pre-generalization aggregation. Therefore, DISTINCT_USERS and TOTAL_PATIENTS are equal for ungeneralized rows (GENERALIZED_ROW = "N"), but TOTAL_PATIENTS will be greater than or equal to DISTINCT_USERS for generalized rows (GENERALIZED_ROW = "Y").

BerryDunn proposes this definition for the TOTAL_PATIENTS field as a useful adjunct to DISTINCT_USERS that provides a more accurate reflection of total service utilization in the commercial market than is reflected in the distinct users count for highly generalized rows.

BerryDunn has provided its programming code to DHHS, NHID, and the departments' data vendor to provide detailed documentation of this methodology and facilitate initial production of the enhanced public use files.

Table 4

Medical, Pharmacy, and Dental Membership by Payer (Files A4-A6)

Row Definition Fields	Order of Generalization	Field Name	Common Name	NH CHIS Data Element ID	Type	Length	Description
✓		service_year	Service Year	MC059, PC032	Number	4	
		coverage_class	Coverage Class	MC899, PC899	String	3	This field indicated the type of record. For all medical claims records, this value will be MED. Pharmacy Claims are PHM. This field is not part of the row definition, as medical, dental, and pharmacy reports are presented separately.
✓	7	parent_payer_code	Parent Payer Code	MC001, PC001	String	8	This field is the Payer ID Number of the data submitter company that links to the REF_PAYER file using the Parent_Payer_Code value. This code is used to identify the data reporter. It is based upon MC001, PC001, or DC001.
		company	Company Name	N/A	String	TBD	Name of the data submitting company associated with the parent_payer_code in the NH Consolidated CHIS REF_PAYER table (e.g., Anthem Health Plans of New Hampshire, Inc.)
✓	8	uf_prim_elig	Primary Eligibility Indicator	ME028	String	1	Y/N field indicating whether eligibility summarized on the row are for members' primary health insurance eligibility ("Y") or not ("N")
✓	6	FI_SI	Fully Insured/Self Insured Indicator	ME029	String	3	This field indicates the type of coverage and is used to distinguish self-funded plans from commercially insured plans as reported in ME029. FI_SI applies a simplifying recoding schema to ME029 **(see Recodes section in the documentation for the mapping)**. Valid values of FI_SI are: FI…commercially insured plan SI…self-funded plan OUM…"Other, Unknown, or Masked:" other plan type, unknown plan type, or value masked by generalization algorithm
✓	4	Prod_Type	Insurance Product Type	N/A			This field identifies standardized product types, and reflects a simplifying recoding schema applied to the "Product_Type" field in the NH Consolidated CHIS (**see Recodes section in the documentation for the mapping**). Valid values are: PPO/POS…Commercial Preferred Provider Organization or Point-of-Service product HMO…Commercial HMO OUM…"Other, Unknown, or Masked:" other product type, unknown product type, or value masked by generalization algorithm
✓	5	Mkt_Seg	Market Segment Code	ME030	String	4	This field indicates the size of the group for group policies, or individual market for individual policies, and applies a simplifying schema to ME030, Market Category Code (**see Recodes section in the documentation for the mapping**). Valid values of Mkt_Seg are: IND…For policies sold and issued directly to individuals, other than those sold on a franchise basis, as defined pursuant to RSA 415:19, or a group conversion policies required pursuant to RSA 415:18 VII (a) GSG…For policies sold and issued directly to employers having between1 and 50 employees GLG1…For policies sold and issued directly to employers having between 51 and 99 employees GLG2…For policies sold and issued directly to employers having 100 or more employees OUM…"Other, Unknown, or Masked:" other market segment, unknown market segment, or value masked by generalization algorithm

Table 4 (cont'd)
Medical, Pharmacy, and Dental Membership by Payer (Files A4-A6)

Row Definition Fields	Order of Generalization	Field Name	Common Name	NH CHIS Data Element ID	Type	Length	Description
✓	3	gender	Member Gender	MC012	String	1	M/F/U
✓	1	age_group_code	Member Age Group Code	ME014, MC059, PC032	Number	1	Numeric code associated with age groupings derived from member age fields in the NH Consolidated CHIS. Valied values are: 1…Ages 0-25 2…Ages 26-64 3…Ages 65+ 999…Unknown or masked by generalization algorithm
✓	2	NH_Res	Member NH Residency Indicator	MC015, PC015	String	1	Indicator variable for member NH residence. Valid values are: 0…Member is not an NH resident 1…Member is an NH resident 999…Unknown or masked by generalization algorithm
✓	2	NH_Region	NH Geographical Area	MC016, PC016	String	TBD	NH geographical region, a value-added field derived from Member County on the NH Consolidated CHIS. Valid values are: 1…Seacoast, Portsmouth, & Dover: Strafford and Rockingham Counties 2…Manchester, Nashua, & Concord: Merrimack and Hillsborough Counties 3…Rural Counties: Belknap, Carroll, Cheshire, Coos, Grafton, & Sullivan 999…Not New Hampshire, unknown, or masked by generalization algorithm
		member_months	Total Member Months	N/A	Number	10	This field identifies the sum of all member months reported by the payer for the membership year. No value < 132 can be reported in this field.
		generalized_row	Generalized Row Indicator	N/A	String	1	Derived field indicating whether the data in the row have been affected by the generalization algorithm ("Y"), or reflect an aggregation of the data as reported in the CHIS eligibility files ("N") without generalization.

4.4 Membership – Medical, Pharmacy, and Dental (Files A4 – A6)

Table 4 above shows the proposed layout of the membership files for medical (File A4), pharmacy (File A5), and dental coverage (File A6). As aggregated data, the files do not comprise personal health information. However, to ensure individual privacy, the methodology follows a modified version of the CMS cell suppression standards; no member months value greater than 0 and less than 132 (the number of member months associated with 11 members continuously eligible for one year) is reported.

The layout and methodology for the three coverage classes is identical.

The file initially sums member months by the following "group-by" fields, as they appear in the base test data:

- Payer, as defined by PARENT_PAYER_CODE. Any medical or pharmacy payer code that did not account for at least 1% of paid claim dollar volume was aggregated into a new payer code, "OTHPAYR = Other or Unknown Payer" before the initial aggregation, so that the same set of payers would be reported in the claims-based and membership files. The same aggregation was made for those dental payers that did not account for at least 1% of total dental membership. The payer company name (COMPANY) is joined to the claim results from the Ref_Payer table.
- Primary Eligibility, as defined by the NH CHIS data field UF_PRIM_ELIG ("Primary Eligibility Use Flag"). The field is coded "Y" if the claims summarized on a given row are associated with the members' primary health insurance eligibility, "N" otherwise.
- FI/SI: The FI_SI field is the result of the simplification algorithm described in the general methodology section for the Consolidated NH CHIS field COV_TYPE.
- Product Type: PROD_TYPE is the result of the simplification algorithm described in the general methodology section for the Consolidated NH CHIS field PRODUCT_TYPE.
- Market Segment: MKT_SEG is the result of the simplification algorithm described in the general methodology section for the Consolidated NH CHIS field COV_TYPE.
- Gender: Member gender, as reported in the NH CHIS field SEX (MC012). Valid values are "F," "M," and "U = Unknown or Masked."
- Age group, as described in the general methodology section.
- NH Resident: Based on values of MEM_STATE (ME016), the NH_RES field takes on the value "1" if the reported member state is New Hampshire, "0" otherwise.
- NH Region, as described in the general methodology section.

Generalized Row, a value-added field created after the generalization process, is included on the final output file. A "Y" in this field indicates that some or all of the records summarized on the row

required at least some generalization to be reportable. An "N" indicates all claim lines summarized on the row were reportable under the values appearing for the constituent rows in the test data.

In testing, less than 1% of member months (0.7% for medical, 0.4% for pharmacy, and 0.8% for dental) required generalization after the initial aggregation. BerryDunn applied the following generalization algorithm to address the unreportable cells. After each of the numbered generalizations, the data were re-summarized by the "group-by" variables and the rows re-checked for adequate cell size:

1. Age group: AGE_GROUP was recoded as "999 = Unknown or Masked" for all rows failing the cell reporting standards after the initial aggregation.
2. Geography (NH residence and region): NH_Region was recoded as "999" and NH_State as "0" for all rows failing the cell suppression standards after age generalization. For all three coverage classes, the geography generalization was by far the most impactful, resolving slightly more than 50% of total member months requiring generalization
3. Gender: Gender was recoded as "U" for the remaining rows with inadequate member months.
4. Product Type: PROD_TYPE was recoded as "OUM = Other, Unknown, or Masked" for the remaining rows with inadequate member months.
5. Market Segment: MKT_SEG was recoded as "OUM = Other, Unknown, or Masked" for the remaining rows with inadequate member months.
6. Funding Type: FI_SI was recoded as "OUM = Other, Unknown, or Masked" for the remaining rows with inadequate member months.
7. Payers: PARENT_PAYER_CODE was recoded as "OTHPAYR = Other Payers" for the remaining rows with inadequate units and patients totals. This generalization pass resolves all remaining cells with less than 132 member months.
8. Primary Eligibility: UF_PRIM_ELIG was recoded as "U = Unknown" for any remaining rows with inadequate member months.

Although this generalization algorithm results in no suppressed volume in testing, *any rows with member months less than 132 (but not equal to 0) remaining at this point would need to be suppressed (excluded) from the published files.* Total member months for suppressed volume should not be reported—by definition, total member months on unreportable fully generalized rows will total less than 132, which is, and will almost certainly remain, immaterial.

The final test files for medical and pharmacy include approximately 1,000 rows. The dental file includes about 600 rows.

BerryDunn has provided its programming code to DHHS, NHID, and the departments' data vendor to provide detailed documentation of this methodology and facilitate initial production of the enhanced public use files.

Table 5
Service Code Expense by Member Demographics (File A7)

Row Definition Fields	Order of Generalization	Field Name	Common Name	NH CHIS Data Element ID	Type	Length	Description	On Current Public Use Files?
✓		service_year	Service Year	MC059	Number	4		Yes
		coverage_class	Coverage Class	MC899, PC899	String	3	This field indicated the type of record. For all medical claims records, this value will be MED. Pharmacy Claims are PHM. This field is not part of the row definition, as medical, dental, and pharmacy reports are presented separately.	Yes
✓	3	gender	Member Gender	MC012	String	1	M/F	Yes
✓	1	age_group_code	Member Age Group Code	ME014, MC059, PC032	Number	1	Numeric code associated with age groupings derived from member age fields in the NH Consolidated CHIS. Valied values are: 1…Ages 0-25 2…Ages 26-64 3…Ages 65+ 999…Unknown or masked by generalization algorithm	Derivable
✓	2	NH_Res	Member NH Residency Indicator	MC015, PC015	String	1	Indicator variable for member NH residence. Valid values are: 0…Member is not an NH resident 1…Member is an NH resident 999…Unknown or masked by generalization algorithm	Yes
✓	2	NH_Region	NH Geographical Area	MC016, PC016	String	TBD	NH geographical region, a value-added field derived from Member County on the NH Consolidated CHIS. Valid values are: 1…Seacoast, Portsmouth, & Dover: Strafford and Rockingham Counties 2…Manchester, Nashua, & Concord: Merrimack and Hillsborough Counties 3…Rural Counties: Belknap, Carroll, Cheshire, Coos, Grafton, & Sullivan 999…Not New Hampshire, unknown, or masked by generalization algorithm	Derivable
✓	4	service_code	Service Code	MC054, MC055	String	20	Revenue code or procedure code. If both a revenue code and a procedure code are reported for a medical claim, the revenue code takes precedence if the type of bill indicates the service is a facility inpatient service. Otherwise (i.e., for facility outpatient or facility-based professional services), the CPT/HCPCS code takes precedence. A service code must have a minimum of 100 paid units and 100 distinct users to be reported. Service codes not meeting this minimum are grouped as "rare, unknown, or aggregated services."	Yes
	4	code_class	Service Code Type	N/A	String	9	Indicates whether the service_code field is populated by a CPT/HCPCS professional services procedure code or a revenue code. Values are: CPT/HCPCS ..... Current Procedural Terminology / Healthcare Common Procedure Coding System (HCPCS) code REV ..... Revenue code	N/A
✓	5	MR_LINE_DESC2	Service Category	N/A	String	50	Service Category; the high-level description for the Milliman HCG code (MR_line, linked through MR_line_case_key on medical claims) associated with the claim line.	Yes

Table 5 (cont'd)
Service Code Expense by Member Demographics (File A7)

Row Definition Fields	Order of Generalization	Field Name	Common Name	NH CHIS Data Element ID	Type	Length	Description	On Current Public Use Files?
		total_billed	Total Billed	MC062, PC035	Decimal	10,2	This field identifies the sum of all amounts billed by providers to the plans in the reporting year. This is a money field containing dollars and cents.	Yes
		total_allowed	Total Allowed	MC063, MC067, MC066, MC065, PC036, PC042, PC041, PC040	Decimal	10,2	This field identifies the sum of all plan and member payments for services rendered to the individual in the reporting year. This is a money field containing dollars and cents.	Derivable
		total_paid	Total Paid	MC063, MC064	Decimal	10,2	This field identifies the sum of all plan payments for services rendered in the	Yes
		total_deduct	Total Deductible Amounts	MC067, PC042	Decimal	10,2	This field identifies the sum of all member deductible payments for services rendered in the reporting year. This is a money field containing dollars and cents.	Yes
		total_coins	Total Coinsurance Amounts	MC066, PC041	Decimal	10,2	This field identifies the sum of all member coinsurance payments (the percentage amount a member must pay for a covered service after any deductible amount is met) for services rendered in the reporting year. This is a money field containing dollars and cents.	Yes
		total_copay	Total Copay Amounts	MC066, PC041	Decimal	10,2	This field identifies the sum of all member copay payments (preset, fixed dollar amounts the member is responsible for, usually on a per-visit/per-service basis, after any deductible amounts are met) for services rendered in the reporting year. This is a money field containing dollars and cents.	Yes
		claim_line_count	Count of Claim Lines Summarized per Row	N/A	Number	10	Number of paid claim line items summarized in the present row. Paid claim line items (SV_STAT = "P") are summed with a value of 1. Reversals (SV_STAT = "R") are summed with a value of -1.	Yes
		distinct_users	Uniique Users	N/A	Number	10	Unique members utilizing paid services summarized on a given row. No value < 11 can be reported in this field.	No
		total_patients	Total Patients	N/A	Number	10	Sum of total patients utilizing paid services summarized on a given row. Total_patients will be duplicative of unique users where a single unique user (as defined by person_key as reported in NH CHIS) has received multiple services summarized on a single generalized row. The value of total_patients should be equal to or greater than the value of distinct_users. No value < 11 can be reported in this field.	No
		generalized_row	Generalized Row Indicator	N/A	String	1	Derived field indicating whether the data in the row have been affected by the generalization algorithm ("Y"), or reflect an aggregation of the data as reported in the CHIS claims files ("N") without generalization.	No

4.5 Medical Expense and Members Served by Service and Member Demographics (File A7)

Table 5 above shows the proposed layout of the medical expense by service and member demographics summary file (File A7). As aggregated data, the file does not comprise personal health information. However, to ensure individual privacy, the methodology follows the CMS cell suppression standards, which prohibit any value greater than 0 and less than 11 to be reported for any measure of services or patients (any amount is considered reportable in dollar fields), by applying generalization or suppression where necessary to avoid the display of too-small claim line or patient counts.

The file initially aggregates all medical claim dollars, units, and patients served by the following "group-by" fields, as they appear in the base test data:

- Gender: Member gender, as reported in the NH CHIS field SEX (MC012). Valid values are "F," "M," and "U = Unknown or Masked."
- Age group, as described in the general methodology section.
- NH Resident: Based on values of MEM_STATE (ME016), the NH_RES field takes on the value "1" if the reported member state is New Hampshire, "0" otherwise.
- NH Region, as described in the general methodology section.
- Service Code. SERVICE_CODE is a value-added field developed by BerryDunn to identify what service was delivered. On a given row, the SERVICE_CODE field may be populated by a CPT4_Code, a HCPCS code, or a revenue code[xv] according to the following logic: for services billed on a UB-04, a CPT4 or HCPCS code reported in the procedure code field (PROC_CODE, MC055) supersedes the reported revenue code (REV_CODE, MC054) for Facility Outpatient or Professional Services utilization types, as defined by the BerryDunn-developed field UTILIZATION_TYPE described in the general methodology section. For other services billed on a UB-04, revenue code supersedes. For all professional claims billed on a CMS-1500, SERVICE_CODE takes the value of the reported CPT4 or HCPCS code.
- Code Class. CODE_CLASS is a BerryDunn-developed field indicating whether SERVICE_CODE is populated by a revenue code (value "REV") or a CPT4 or HCPCS code (value "CPT/HCPCS").
- MR_Line_Desc2. The service class description MR_LINE_DESC_2 from the Ref_HCG table is joined to the claims before the initial aggregation by values of Milliman's HCG field

[xv] No service description is included in the design because the SERVICE_CODE field is an amalgam of three distinct code sets.

MR_LINE_CASE_KEY, which appears on both the claims and Ref_HCG. MR_LINE_DESC_2 provides information on the type of service rendered for lines where the service code must be generalized to meet cell reporting guidelines. Please note that MR_LINE_DESC_2 and SERVICE_CODE have a many-to-many relationship.[xvi]

Generalized Row, a value-added field created after the generalization process, is included on the final output file. A "Y" in this field indicates that some or all of the records summarized on the row required at least some generalization to be reportable. An "N" indicates all claim lines summarized on the row were reportable under the values appearing for the constituent rows in the test data. All of the other information included in this file can be derived from the current NH CHIS public use data sets, except the patient counts.

In testing, claims representing $168 million (7%) of allowed volume summed to unreportable rows after the initial aggregation. BerryDunn applied the following generalization algorithm to address the unreportable cells. After each of the numbered generalizations, the data were re-summarized by the "group-by" variables and the rows re-checked for adequate cell size:

1. Age group: AGE_GROUP was recoded as "999 = Unknown or Masked" for all rows failing the cell reporting standards after the initial aggregation. This generalization rendered 30% ($50 million allowed) of the unreportable volume reportable, leaving $119 million allowed (5% of commercial medical claims total) that required further generalization.
2. Geography (NH residence and region): NH_Region was recoded as "999" and NH_State as "0" for all rows failing the cell suppression standards after age generalization. This generalization rendered 75% ($89 million allowed) of the remaining unreportable volume reportable, leaving $29 million (1% of commercial medical claims total) that required further generalization.
3. Gender: Gender was recoded as "U" for the remaining rows with inadequate units and patients totals. This generalization resolved 29% ($8 million allowed) of the remaining unreportable volume, leaving $21 million (1% of commercial medical claims total) that required further generalization.
4. Service Code: SERVICE_CODE was recoded as "AGGRGTDSVC = Aggregated Services" and CODE_CLASS recoded as "OUM = Other, Unknown, or Masked" for the remaining rows with inadequate units and patients totals. This generalization rendered 99.98% ($21 million allowed) of the unreportable volume reportable, leaving an immaterial $4,495 allowed that required further generalization.

[xvi] However, this does not result in double counting because each claim line is associated with exactly one SERVICE_CODE and MR_LINE_DESC_2, and the join of MR_LINE_CASE_KEY between claims and Ref_HCG is many-to-one.

5. MR_Line_Desc2: MR_LINE_DESC_2 was recoded as "Generalized" for the remaining rows failing the cell suppression standards, and the aggregation was repeated. This generalization rendered all remaining rows reportable.

Although this generalization algorithm results in no suppressed volume in testing, any rows with users or claim lines less than 11 (but not equal to 0) remaining at this point would need to be suppressed (excluded) from the published files.

A summary of dollars for suppressed volume should be provided in the companion documentation if suppression is required in production.

Similar summaries of suppressed claim line counts and users should *not* be reported. If the totals for these fields do not reach the reportability threshold, such a summary would fail the reportability standards; if, on the other hand, the totals are reported where they do meet the threshold, their absence in the companion documentation to other files (where the threshold was not met) would be tantamount to reporting totals beneath the threshold.

The final test file includes approximately 66,000 rows.

Two distinct concepts of service user counts are included in the aggregated files. DISTINCT_USERS provides a count of distinct values of PERSON_KEY represented on all claim lines summarized on the row, regardless of the level of generalization.

The TOTAL_PATIENTS field, however, counts all users for all key variable ("group-by") combinations summarized on the row. That is, if the same member receives services under two distinct service codes reported on rows requiring service code generalization, those rows will add 1 to the DISTINCT_USERS count, and 2 to the TOTAL_PATIENTS count.

TOTAL_PATIENTS is calculated by summing the value of DISTINCT_USERS calculated in the initial, pre-generalization aggregation. Therefore, DISTINCT_USERS and TOTAL_PATIENTS are equal for ungeneralized rows (GENERALIZED_ROW = "N"), but TOTAL_PATIENTS will be greater than or equal to DISTINCT_USERS for generalized rows (GENERALIZED_ROW = "Y").

BerryDunn proposes this definition for the TOTAL_PATIENTS field as a useful adjunct to DISTINCT_USERS that provides a more accurate reflection of total service utilization in the commercial market than is reflected in the distinct users count for highly generalized rows.

BerryDunn has provided its programming code to DHHS, NHID, and the departments' data vendor to provide detailed documentation of this methodology and facilitate initial production of the enhanced public use files.

Table 6

Medical Expenses by Three-Digit Primary Diagnosis and Member Demographics (File A8)

Row Definition Fields	Order of Generalization	Field Name	Common Name	NH CHIS Data Element ID	Type	Length	Description	On Current Public Use Files?
✓		service_year	Service Year	MC059	Number	4		Yes
		coverage_class	Coverage Class	MC899, PC899	String	3	This field indicated the type of record. For all medical claims records, this value will be MED. Pharmacy Claims are PHM. This field is not part of the row definition, as medical, dental, and pharmacy reports are presented separately.	Yes
✓	3	gender	Member Gender	MC012	String	1	M/F	Yes
✓	1	age_group_code	Member Age Group Code	ME014, MC059, PC032	Number	1	Numeric code associated with age groupings derived from member age fields in the NH Consolidated CHIS. Valied values are: 1…Ages 0-25 2…Ages 26-64 3…Ages 65+ 999…Unknown or masked by generalization algorithm	Derivable
✓	2	NH_Res	Member NH Residency Indicator	MC015, PC015	String	1	Indicator variable for member NH residence. Valid values are: 0…Member is not an NH resident 1…Member is an NH resident 999…Unknown or masked by generalization algorithm	Yes
✓	2	NH_Region	NH Geographical Area	MC016, PC016	String	TBD	NH geographical region, a value-added field derived from Member County on the NH	Derivable
✓	4	dx1_code	Principal Diagnosis Code	MC041	String	3	First three digits of the principal ICD diagnosis code.	Derivable
	4	icd10_desc	Diagnosis Description	N/A	String		ICD-10 "long description" for dx1_code. Only populated for ICD-10 values (ICD-10 cutover was October 1, 2015).	No
	5	icd10_chapter	ICD-10 Chapter	N/A	Number	2	ICD-10 chapter number of the 3-digit ICD-10 code. Only populated for ICD-10 codes.	No
	5	icd10_chapter_name	ICD-10 Chapter Name	N/A	String		Name of the ICD-10 Chapter (e.g., Chapter 13, "Diseases of the Musculoskeletal System and Connective Tissue (M00-M99)").	No

Table 6 (cont'd)

Medical Expenses by Three-Digit Primary Diagnosis and Member Demographics (File A8)

Row Definition Fields	Order of Generalization	Field Name	Common Name	NH CHIS Data Element ID	Type	Length	Description	On Current Public Use Files?
		total_billed	Total billed	MC062, PC035	Decimal	10,2	This field identifies the sum of all amounts billed by providers to the plans in the reporting year. This is a money field containing dollars and cents.	Yes
		total_allowed	Total Allowed	MC063, MC067, MC066, MC065, PC036, PC042, PC041, PC040	Decimal	10,2	This field identifies the sum of all plan and member payments for services rendered to the individual in the reporting year. This is a money field containing dollars and cents.	Derivable
		total_paid	Total Paid	MC063, MC064	Decimal	10,2	This field identifies the sum of all plan payments for services rendered in the reporting year. This is a money field containing dollars and cents.	Yes
		total_deduct	Total Deductible Amounts	MC067, PC042	Decimal	10,2	This field identifies the sum of all member deductible payments for services rendered in the reporting year. This is a money field containing dollars and cents.	Yes
		total_coins	Total Coinsurance Amounts	MC066, PC041	Decimal	10,2	This field identifies the sum of all member coinsurance payments (the percentage amount a member must pay for a covered service after any deductible amount is met) for services rendered in the reporting year. This is a money field containing dollars and cents.	Yes
		total_copay	Total Copay Amounts	MC066, PC041	Decimal	10,2	This field identifies the sum of all member copay payments (preset, fixed dollar amounts the member is responsible for, usually on a per-visit/per-service basis, after any deductible amounts are met) for services rendered in the reporting year. This is a money field containing dollars and cents.	Yes
		claim_line_count	Count of Claim Lines Summarized per Row	N/A	Number	10	Number of paid claim line items summarized in the present row. Paid claim line items (SV_STAT = "P") are summed with a value of 1. Reversals (SV_STAT = "R") are summed with a value of -1.	Yes
		distinct_users	Uniique Users	N/A	Number	10	Unique members utilizing paid services summarized on a given row. No value < 11 can be reported in this field.	No
		total_patients	Total Patients	N/A	Number	10	Sum of total patients utilizing paid services summarized on a given row. Total_patients will be duplicative of unique users where a single unique user (as defined by person_key as reported in NH CHIS) has received multiple services summarized on a single generalized row. The value of total_patients should be equal to or greater than the value of distinct_users. No value < 11 can be reported in this field.	No
		generalized_row	Generalized Row Indicator	N/A	String	1	Derived field indicating whether the data in the row have been affected by the generalization algorithm ("Y"), or reflect an aggregation of the data as reported in the CHIS claims files ("N") without generalization.	No

4.6 Medical Expense and Total Patients by Primary Diagnosis and Member Demographics (File A8)

Table 6 above shows the proposed layout of the medical expense by three-digit primary diagnosis and member demographics summary file (File A8). As aggregated data, the file does not comprise personal health information. However, to ensure individual privacy, the methodology follows the CMS cell suppression standards, which prohibit any value greater than 0 and less than 11 to be reported for any measure of services or patients (any amount is considered reportable in dollar fields), by applying generalization or suppression where necessary to avoid the display of too-small claim or patient counts.

The file initially aggregates all medical claim dollars, claim line counts, and patients served by the following "group-by" fields, as they appear in the base test data:

- Gender: Member gender, as reported in the NH CHIS field SEX (MC012). Valid values are "F," "M," and "U = Unknown or Masked."
- Age group, as described in the general methodology section.
- NH Resident: Based on values of MEM_STATE (ME016), the NH_RES field takes on the value "1" if the reported member state is New Hampshire, "0" otherwise.
- NH Region, as described in the general methodology section.
- Three-Digit Principal Diagnosis Code: the DX1_CODE field in the proposed design shows the values of the first three digits of the principal diagnosis (ICD_DIAG_01_PRIMARY, MC041) reported on the claim. Three-digit diagnosis codes appearing for fewer than 25 distinct individuals were generalized under the code "GEN" prior to the initial aggregation. This generalization affected claims representing approximately $12 million allowed (0.5% of total).
- Three-Digit Diagnosis Code Description: descriptions for ICD-10 three-digit diagnosis codes were joined to the claims from the ICD-10 electronic reference files published by the American Medical Association.
- ICD-10 Chapter Number and Name: the ICD-10 chapter number and name for each ICD-10 code, derived from the ICD-10 reference manual published by the American Medical Association.

Generalized Row, a value-added field created after the generalization process, is included on the final output file. A "Y" in this field indicates that some or all of the records summarized on the row required at least some generalization to be reportable. An "N" indicates all claim lines summarized on the row were reportable under the values appearing for the constituent rows in the test data. All of the other information included in this file can be derived from the current CHIS public use data sets and publicly-available sources, except the patient counts.

In testing, claims representing $115 million (5%) of allowed volume summed to unreportable rows (rows with less than 11 units or users) after the initial aggregation. BerryDunn applied the following generalization algorithm to address the unreportable cells. After each of the numbered generalizations, the data were re-summarized by the "group-by" variables and the rows re-checked for adequate cell size:

1. Age group: AGE_GROUP was recoded as "999 = Unknown or Masked" for all rows failing the cell reporting standards after the initial aggregation. This generalization rendered 38% ($44 million allowed) of the unreportable volume reportable, leaving $71 million allowed (3% of commercial medical claims total) that required further generalization.
2. Geography (NH residence and region): NH_Region was recoded as "999" and NH_State as "0" for all rows failing the cell suppression standards after age generalization. This generalization rendered 87% ($62 million allowed) of the remaining unreportable volume reportable, leaving $9 million (0.4% of commercial medical claims total) that required further generalization.
3. Gender: Gender was recoded as "U" for the remaining rows with inadequate claim line and patient totals. This generalization resolved 22% ($1.9 million allowed) of the remaining unreportable volume, leaving $7 million (0.3% of commercial medical claims total) that required further generalization.
4. Three-digit diagnosis and description: the three-digit diagnosis code was recoded to "GEN" and the description to "Generalized" for the remaining rows with inadequate claim line and patient totals. This generalization resolved all rows but for rows representing an immaterial $419 in allowed charges.
5. Chapter: ICD10_CHAPTER was generalized to "99" and ICD10_CHAPTER_NAME was generalized to "Other, Unknown, or Masked" for the remaining rows with inadequate claim line and patient totals. This final generalization pass did not resolve the remaining unreportable rows.

The remaining rows with users or claim lines less than 11 (but not equal to 0) remaining at this point would need to be suppressed (excluded) from the published files.

A summary of dollars for suppressed volume should be provided in the companion documentation if suppression is required in production.

Similar summaries of suppressed claim line counts and users should *not* be reported. If the totals for these fields do not reach the reportability threshold, such a summary would fail the reportability standards; if, on the other hand, the totals are reported where they do meet the threshold, their absence in the companion documentation to other files (where the threshold was not met) would be tantamount to reporting totals beneath the threshold.

The final test file includes approximately 16,000 rows.

Two distinct concepts of service user counts are included in the aggregated files. DISTINCT_USERS provides a count of distinct values of PERSON_KEY represented on all claim lines summarized on the row, regardless of the level of generalization.

The TOTAL_PATIENTS field, however, counts all users for all key variable ("group-by") combinations summarized on the row. That is, if the same member receives services with two distinct three-digit principal diagnoses reported on rows requiring diagnosis code generalization, those rows will add 1 to the DISTINCT_USERS count, and 2 to the TOTAL_PATIENTS count.

TOTAL_PATIENTS is calculated by summing the value of DISTINCT_USERS calculated in the initial, pre-generalization aggregation. Therefore, DISTINCT_USERS and TOTAL_PATIENTS are equal for ungeneralized rows (GENERALIZED_ROW = "N"), but TOTAL_PATIENTS will be greater than or equal to DISTINCT_USERS for generalized rows (GENERALIZED_ROW = "Y").

BerryDunn proposes this definition for the TOTAL_PATIENTS field as a useful adjunct to DISTINCT_USERS that provides a more accurate reflection of total service utilization in the commercial market than is reflected in the distinct users count for highly generalized rows.

BerryDunn has provided its programming code to DHHS, NHID, and the departments' data vendor to provide detailed documentation of this methodology and facilitate initial production of the enhanced public use files.

Table 7

Pharmacy Expenses by Drug and Member Demographics (File A9)

Row Definition Fields	Order of Generalization	Field Name	Common Name	NH CHIS Data Element ID	Type	Length	Description
✓		service_year	Service Year	PC032	Number	4	
		coverage_class	Coverage Class	MC899, PC899	String	3	This field indicated the type of record. For all medical claims records, this value will be MED. Pharmacy Claims are PHM. This field is not part of the row definition, as medical, dental, and pharmacy reports are presented separately.
✓	3	gender	Member Gender	PC012	String	1	M/F
✓	1	age_group_code	Member Age Group Code	ME014, MC059, PC032	Number	1	Numeric code associated with age groupings derived from member age fields in the NH Consolidated CHIS. Valied values are: 1…Ages 0-25 2…Ages 26-64 3…Ages 65+ 999…Unknown or masked by generalization algorithm
✓	2	NH_Res	Member NH Residency Indicator	MC015, PC015	String	1	Indicator variable for member NH residence. Valid values are: 0…Member is not an NH resident 1…Member is an NH resident 999…Unknown or masked by generalization algorithm
✓	2	NH_Region	NH Geographical Area	MC016, PC016	String	TBD	NH geographical region, a value-added field derived from Member County on the NH Consolidated CHIS. Valid values are: 1…Seacoast, Portsmouth, & Dover: Strafford and Rockingham Counties 2…Manchester, Nashua, & Concord: Merrimack and Hillsborough Counties 3…Rural Counties: Belknap, Carroll, Cheshire, Coos, Grafton, & Sullivan 999…Not New Hampshire, unknown, or masked by generalization algorithm
✓	4	NDC	Drug NDC Code	PC026	String	11	FDA national drug code of the dispensed drug.
✓	5	ndc_prod_name	Drug Name	N/A	String	50	Drug name, as reported by the carrier in the NH CHIS pharmacy claims file.
✓	5	brand_status	Generic or Brand indicator	N/A	String	7	This field is recoded from Brand_Status in NH CHIS (**see Recodes section of the documentation for the mapping**): GENERIC….Generic drug BRAND…Brand-name Drug OUM…Other, Unknown, or Masked
✓	6	drug_ther_class	Drug Therapeutic Class Description	N/A	String		Description of the therapeutic class code (NDC_ther_class) field reported on the pharmacy claim line

Table 7 (cont'd)
Pharmacy Expenses by Drug and Member Demographics (File A9)

Row Definition Fields	Order of Generalization	Field Name	Common Name	NH CHIS Data Element ID	Type	Length	Description
		total_billed	Total billed	MC062, PC035	Decimal	10,2	This field identifies the sum of all amounts billed by providers to the plans in the reporting year. This is a money field containing dollars and cents.
		total_allowed	Total Allowed	MC063, MC067, MC066, MC065, PC036, PC042, PC041, PC040	Decimal	10,2	This field identifies the sum of all plan and member payments for services rendered to the individual in the reporting year. This is a money field containing dollars and cents.
		total_paid	Total Paid	PC036	Decimal	10,2	This field identifies the sum of all plan payments for services rendered in the reporting year. This is a money field containing dollars and cents.
		total_deduct	Total Deductible Amounts	MC067, PC042	Decimal	10,2	This field identifies the sum of all member deductible payments for services rendered in the reporting year. This is a money field containing dollars and cents.
		total_coins	Total Coinsurance Amounts	MC066, PC041	Decimal	10,2	This field identifies the sum of all member coinsurance payments (the percentage amount a member must pay for a covered service after any deductible amount is met) for services rendered in the reporting year. This is a money field containing dollars and cents.
		total_copay	Total Copay Amounts	MC066, PC041	Decimal	10,2	This field identifies the sum of all member copay payments (preset, fixed dollar amounts the member is responsible for, usually on a per-visit/per-service basis, after any deductible amounts are met) for services rendered in the reporting year. This is a money field containing dollars and cents.
		total_days_supply	Total days' supply	PC034	Decimal	10,2	This field identifies the total days' supply of paid utilization. No absolute value < 11 can be reported in this field.
		claim_line_count	Count of Claim Lines Summarized per Row	N/A	Number	10	Number of paid claim line items summarized in the present row. Paid claim line items (SV_STAT = "P") are summed with a value of 1. Reversals (SV_STAT = "R") are summed with a value of -1.
		distinct_users	Uniique Users	N/A	Number	10	Unique members utilizing paid services summarized on a given row. No value < 11 can be reported in this field.
		total_patients	Total Patients	N/A	Number	10	Sum of total patients utilizing paid services summarized on a given row. Total_patients will be duplicative of unique users where a single unique user (as defined by person_key as reported in NH CHIS) has received multiple services summarized on a single generalized row. The value of total_patients should be equal to or greater than the value of distinct_users. No value < 11 can be reported in this field.
		generalized_row	Generalized Row Indicator	N/A	String	1	Derived field indicating whether the data in the row have been affected by the generalization algorithm ("Y"), or reflect an aggregation of the data as reported in the CHIS claims files ("N") without generalization.

4.7 Pharmacy Expense and Total Patients by Drug (File A9)

Table 7 above shows the proposed layout of the pharmacy expense by drug and member demographics summary file (File A9). As aggregated data, the file does not comprise personal health information. However, to ensure individual privacy, the methodology follows the CMS cell suppression standards, which prohibit any value greater than 0 and less than 11 to be reported for any measure of services or patients (any amount is considered reportable in dollar fields), by applying generalization or suppression where necessary to avoid the display of too-small unit or patient counts.

The file initially aggregates all pharmacy claim dollars, units, and patients served by the following "group-by" fields, as they appear in the base test data (except as noted):

- Gender: Member gender, as reported in the NH CHIS field SEX (MC012). Valid values are "F," "M," and "U = Unknown or Masked."
- Age group, as described in the general methodology section.
- NH Resident: Based on values of MEM_STATE (ME016), the NH_RES field takes on the value "1" if the reported member state is New Hampshire, "0" otherwise.
- NH Region, as described in the general methodology section.
- National Drug Code: the NDC field shows the values of National Drug Code (NDC, PC026) reported on the claim. Drugs appearing for fewer than 25 distinct individuals or less than 25 days supply were generalized along with any invalid NDCs under the code "RAREORUNKRX" prior to the initial aggregation. This generalization affected claims representing approximately $114 million allowed (approximately 12% of total).
- Drug Name: the DRUG_NAME field displays the value of the drug name (NDC_PROD_NAME) as reported on the claims. Drug name was generalized to "RARE OR UNKNOWN DRUG" prior to generalization for all rows affected by the NDC drug code generalization described immediately above.
- Brand Status: the BRAND_STATUS field indicates whether a prescription was reported as filled by a generic or brand name product. As it appears in the present file, it takes on the values "BRAND," "GENERIC," and "OUM" as described in the simplification algorithm in the general methodology section, above.
- Drug Therapeutic Class: DRUG_THER_CLASS shows the values of the MAIN field in the Ref_NDC_Ther_Class table, merged onto the pharmacy claims by the field NDC_THER_CLASS on pharmacy claims joined to THER_CLASS on Ref_NDC_Ther_Class.

Generalized Row, a value-added field created after the generalization process, is included on the final output file. A "Y" in this field indicates that some or all of the records summarized on the row required at least some generalization to be reportable. An "N" indicates all claim lines summarized

on the row were reportable under the values appearing for the constituent rows in the test data. All of the other information included in this file can be derived from the current CHIS public use data sets, except the patient counts.

Claims representing $220 million (22%) of allowed volume summed to unreportable rows (rows with less than 11 claim lines, days supply, or users) after the initial aggregation. BerryDunn applied the following generalization algorithm to address the unreportable cells. After each of the numbered generalizations, the data were re-summarized by the "group-by" variables and the rows re-checked for adequate cell size:

1. Age group: AGE_GROUP was recoded as "999 = Unknown or Masked" for all rows failing the cell reporting standards after the initial aggregation. This generalization rendered 23% ($51 million allowed) of the unreportable volume reportable, leaving $170 million allowed (17% of commercial pharmacy claims total) that required further generalization.
2. Geography (NH residence and region): NH_Region was recoded as "999" and NH_State as "0" for all rows failing the cell suppression standards after age generalization. This generalization rendered 88% ($150 million allowed) of the remaining unreportable volume reportable, leaving $20 million (2% of commercial pharmacy claims total) that required further generalization.
3. Gender: Gender was recoded as "U" for the remaining rows with inadequate claim line, days supply, or patient totals. This generalization resolved 22% ($4.4 million allowed) of the remaining unreportable volume, leaving $15.5 million (1.5% of commercial pharmacy claims total) that required further generalization.
4. NDC: NDC was recoded as "RAREORUNKRX" for the remaining rows with inadequate claim line, days supply, or patient totals. This generalization resolved 29% ($4.5 million allowed) of the remaining unreportable volume, leaving $11 million (1.1% of commercial pharmacy claims total) that required further generalization.
5. Drug Name and Brand Status: multiple NDCs often map to the same drug name, allowing drug name to remain reportable for some rows requiring NDC generalization. For those rows remaining unreportable after NDC generalization, drug name was recoded as "RARE OR UNKNOWN DRUG," and brand status was recoded as "OUM = Other, Unknown, or Masked." This generalization resolved 97% ($10.7 million allowed) of the remaining unreportable volume, leaving $376,000 (0.04% of commercial pharmacy claims total) that required further generalization.
6. Therapeutic Class: recoding DRUG_THER_CLASS to "GENERALIZED" resolved all remaining unreportable rows.

Although this generalization algorithm results in no suppressed volume in testing, *any rows with users, days supply, or claim lines less than 11 (but not equal to 0) remaining at this point would need to be suppressed (excluded) from the published files.*

A summary of dollars for suppressed volume should be provided in the companion documentation if suppression is required in production.

Similar summaries of suppressed claim line counts, days supply, and users should *not* be reported. If the totals for these fields do not reach the reportability threshold, such a summary would fail the reportability standards; if, on the other hand, the totals are reported where they do meet the threshold, their absence in the companion documentation to other files (where the threshold was not met) would be tantamount to reporting totals beneath the threshold.

The final test file includes approximately 68,000 rows.

Two distinct concepts of service user counts are included in the aggregated files. DISTINCT_USERS provides a count of distinct values of PERSON_KEY represented on all claim lines summarized on the row, regardless of the level of generalization.

The TOTAL_PATIENTS field, however, counts all users for all key variable ("group-by") combinations summarized on the row. That is, if the same member fills prescriptions in two different therapeutic classes reported on rows requiring therapeutic class generalization, those rows will add 1 to the DISTINCT_USERS count, and 2 to the TOTAL_PATIENTS count.

TOTAL_PATIENTS is calculated by summing the value of DISTINCT_USERS calculated in the initial, pre-generalization aggregation. Therefore, DISTINCT_USERS and TOTAL_PATIENTS are equal for ungeneralized rows (GENERALIZED_ROW = "N"), but TOTAL_PATIENTS will be greater than or equal to DISTINCT_USERS for generalized rows (GENERALIZED_ROW = "Y").

BerryDunn proposes this definition for the TOTAL_PATIENTS field as a useful adjunct to DISTINCT_USERS that provides a more accurate reflection of total service utilization in the commercial market than is reflected in the distinct users count for highly generalized rows.

BerryDunn has provided its programming code to DHHS, NHID, and the departments' data vendor to provide detailed documentation of this methodology and facilitate initial production of the enhanced public use files.

Table 8
Population Cost by Member – Medical (File P1)

Row Definition Fields	Order of Generalization	Field Name	Common Name	NH CHIS Data Element ID	Type	Length	Description	On Current Public Use Files?
✓		service_year	Service Year	MC059, PC032	Number	4		Yes
✓		person_key	Hashed person key	N/A	TBD	TBD	Must be "re-hashed" for every release to avoid intertemporal linkage	No
		coverage_class	Coverage Class	MC899	String	3	This field indicated the type of record. For all medical claims records, this value will be MED. Pharmacy Claims are PHM.	Yes
✓	2	CCHG_Cat	Chronic Condition Hierarchical Grouping Code	N/A	String	4	Chronic Condition Hierarchical Groups organize medical utilization and costs in a clinically relevant manner. Milliman's CCHGs assign patients to unique categories using a clinically relevant hierarchy that groups similar patients in the same group based on how doctors make treatment decisions. i.e. 101 - Major Psychosis, 114 - COPD, or 130 - Healthy Male (41-64). The population cost by member file reflects the following changes to Milliman's CCHG_Cat assignment: -The ten "Healthy Male" and "Healthy Female" age group categories for ages 65 and over have been recoded to the categories "145...Healthy Male (65+)" and "146...Healthy Female (65+)" -All claims for a member in the year are assigned the member's most recent valid CCHG_Cat (as defined by claim line service from date), if any. -Null CCHG_Cat values and CCHG_Cat values masked to achieve adequate equivalence class size (*k*>=20) are set to "999...Unknown or Masked CCHG"	Yes
		CCHG_Cat_Desc	Chronic Condition Hierarchical Grouping Description	N/A	String	92	Description of the CCHG code (e.g., "Major Psychosis," "Healthy Female (6-15)").	Yes
✓	1	Utilization_Category	Utilization Category	MC036, MC037	String	60	This field identifies the utilization category for this record. It is a value-added field based on the Type of Bill field for institutional claims and Place of Service for professional claims. Valid values are: Ambulatory Surgery Clinic/Office Home Health/Hospice Hospital Inpatient Hospital Inpatient Professional Services Hospital Outpatient Nursing Facility OUM	Derivable
		total_allowed	Total Allowed	MC063, MC067, MC066, MC065, PC036, PC042, PC041, PC040	Decimal	10,2	This field identifies the sum of all plan and member payments for services rendered to the individual in the reporting year. This is a money field containing dollars and cents.	Not Derivable at Person Level
		total_paid	Total Paid	MC063, PC036, MC064	Decimal	10,2	This field identifies the sum of all plan payments for services rendered to the individual in the reporting year. This is a money field containing dollars and cents.	Not Derivable at Person Level
		claim_line_count	Count of Claim Lines Summarized per Row	N/A	Number	10	Number of paid claim line items summarized in the present row. Paid claim line items (SV_STAT = "P") are summed with a value of 1. Reversals (SV_STAT = "R") are summed with a value of -1.	Not Derivable at Person Level

5 De-Identified Person-Level Files

The proposed design includes two files of de-identified person-level health data:

- Population Medical Cost by Member (File P1)
- Population Pharmacy Cost by Member (File P2)

The following two subsections describe the designs, calculation, de-identification methodology, and results for each file.

5.1 Population Medical Cost by Member File (File P1)

Table 8 above shows the proposed layout of the population medical cost by member file (File P1). The file shows total commercial allowed charges, carrier paid amounts (including prepaid amounts), and claim line counts summarized by individual,[xvii] general diagnostic category, and utilization type. The presence of an encrypted person-level identifier (the PERSON_KEY field) in the file renders the content personal information, and re-identification risks must be managed as such.

The academic literature on the de-identification of personal health information suggests a maximum re-identifiability risk of 5% for a public data set's most vulnerable record.[28]

As noted above, this standard requires removal or encryption of any direct identifiers and the creation of adequate equivalence classes. A simple metric for assessing re-identification risk, or the probability of re-identification (Pr(ID)), posits that re-identification risk is no greater than the inverse of the size of the minimum equivalence class; that is, given that a probability of re-identification less than or equal to 5% is equivalent to:

$$\Pr(\mathrm{ID}) \leq 1/20$$

requiring equivalence classes of 20 for all combinations of quasi-identifiers in a population data set and effectively encrypting direct identifiers meets a generally accepted standard of de-identification of person-level health care data for public use.

[xvii] In the report designs, a distinct PERSON_KEY defines a unique individual. Throughout this discussion, the terms "PERSON_KEY," "member," and "individual" are used interchangeably. The Consolidated NH CHIS documentation defines this field as "the key that identifies a unique person within the data warehouse." If a single value of PERSON_KEY appears under multiple payer codes, the volume for all payers associated with the PERSON_KEY value are summed together on the lines reporting that PERSON_KEY. On the other hand, if a person's claims or eligibility are reported under multiple PERSON_KEYs in the NH CHIS medical data, that person's data will be reported under multiple PERSON_KEYs in the population cost file, as if the records belonged to multiple people.

Testing the sizes of equivalence classes is often referred to in the literature as "testing *k*-anonymity," where *k* is the number of members required to form an adequately de-identified equivalence class.

File P1 includes two quasi-identifiers:

- Chronic Condition Hierarchical Grouping Category (CCHG_CAT) and description: these categories are value-added data developed by DHHS's NH CHIS data vendor that "organize medical utilization and costs in a clinically relevant manner. [The] CCHGs assign patients to unique categories using a clinically relevant hierarchy that groups similar patients in the same group based on how doctors make treatment decisions. i.e. 101 – Major Psychosis, 114 – COPD, or 130 – Healthy Male (41 – 64)."[29] In the test file, 80% of PERSON_KEYs have a unique Chronic Condition Hierarchical Grouping for the year. For those PERSON_KEYs with multiple Chronic Condition Hierarchical Groupings reported, BerryDunn assigned a single, unique Chronic Condition Hierarchical Grouping to all claims for a single PERSON_KEY by assigning all claims to the last valid Chronic Condition Hierarchical Grouping reported in the year by service date.[xviii]
- Utilization Type (UTILIZATION_TYPE), a value-added field developed by BerryDunn as described above in the "Test Data" section.

Age group, gender, and member geography quasi-identifiers were tested in this file but removed because:

- Including these fields exponentially increases the claim volume requiring generalization to achieve adequate equivalence classes, and
- Chronic Condition Hierarchical Grouping identifies age and gender (e.g., "138 = Healthy Female (41 – 64)") for approximately 50% of individuals in the file.

The following subsections detail the risk-management steps required to produce the de-identified medical expense by member file.

Re-encryption of PERSON_KEY

As noted above, file P1 includes the Consolidated NH CHIS member identifier PERSON_KEY. Although PERSON_KEY, as it appears in the Consolidated NH CHIS, is already encrypted and is therefore not a direct member identifier, the enhanced public use file methodology requires the

[xviii] That is, if a PERSON_KEY had a single claim each in January, September, and December 2016, if the January 2016 claim reported a CCHG_CAT of "125 = Other chronic conditions," the September 2016 claim reported "130 = Healthy Male (41 – 64), and the December claim reported a null CCHG_CAT, CCHG_CAT "130 = Healthy Male (41 – 64) would be assigned to all three claims in this file.

following re-encryption to minimize the risk of member re-identification by linking between NH CHIS data sources:

1. All values of PERSON_KEY must be randomly assigned a new value (rather than perturbing each value by a simple transformation, or "hash"). DHHS's data vendor should securely maintain a mapping of the Consolidated NH CHIS PERSON_KEY values to the population cost by member files' values.
2. PERSON_KEY values for the medical and pharmacy versions of the population cost by member files should be reassigned separately, such that an individual's data may *not* be linked across the two files.
3. Each year's file should be encrypted separately such that an individual's data may *not* be linked across files for multiple years.

Numbers 2 and 3 above are required because *k*-anonymity is tested in isolation for the data in each coverage class (medical or pharmacy) and each year. The methodology does not require that the data reported in the 2016 medical file for a given individual form adequate equivalence classes when combined with data for that individual in the 2016 pharmacy file, or in the 2017 medical file.

Random Sampling

The first step in producing the file is to identify a 50% random sample of members with paid utilization and limit the analysis data set to claim lines associated with only the members in the sample. This choice has two main advantages for the present purpose. First, and most important, sampling provides a measure of security against attempts to re-identify observations in the person-level files by linking to additional data in other data sources (particularly other files in the enhanced NH CHIS public use files design). Second, when a sample of personal data are taken prior to de-identification, the equivalence class size required for de-identification can be reduced in proportion to the size of the sample.[30]

After the 50% random sample data set is created, total expenses, claim lines, and users in the sample file should be checked against the population totals to ascertain that the random sampling process has yielded a data set including half the volume of the population file on those aggregated measures. BerryDunn recommends the sample be tested against a standard of including 50% of total allowed charges, payments, claim counts, and distinct users when the ratio of the totals in the sample to the totals in the population file are rounded to the 1% (hundredth) place.

BerryDunn's resulting medical claims 50% sample test file included:

- 6,217,575 claim lines, less reversals (50% of 12,464,227 claim lines in the full test file)
- $1,187,121,149 allowed (50% of $2,374,869,392)
- $943,378,355 paid, including prepaid amounts (50% of $1,889,306,120)

- 276,735 distinct users (50% of 553,543)

The following subsections describe the advantageous consequences of sampling in detail.

Sampling protects against file linkage

Sampling provides security against attempts to re-identify observations in the person-level files by linking to additional data in other data sources. Specifically, the person-level medical expense file and the medical expense by payer and product type aggregated file (File A2) both summarize spending by the value-added field UTILIZATION_CATEGORY, among others. In theory, an adversary could search the person-level medical expense file for sets of members whose quantitative data (allowed and paid dollar totals and claim line counts) for a given utilization category row match a row in file A2. Finding such a match in a population person-level file, the adversary would then have associated all the quasi-identifiers in file A2 with all the records for all utilization categories associated with those members in file P1.

For example, consider the following fictional example record from the medical expense by payer and product type file:

Parent Payer Code	Company	FI/ SI	Prod. Type	Mkt. Seg.	Utilization Type	Gender	Age Group	NH Res.	NH Region	UF Prim. Elig.	Total Allowed	Total Paid	Claim Line Count	Total Patients	Distinct Users	Generalized Row
NHC999	Acme Insurance	FI	PPO/POS	IND	Nursing Facility	F	2	1	3	Y	$65,432	$54,321	42	11	11	N

The adversary could search all Nursing Facility records in file P1 for a unique set of 11 members with exactly 42 claim lines, $65,432 allowed, and $54,321 paid among them. Finding such a match, the adversary would then have associated a carrier, funding source, product type, market segment, gender, age group, New Hampshire region of residence, and primary eligibility indicator with all the diagnostic category and utilization type records in file P1 for those 11 members. For example, suppose one of the members in the matching set had the following records in the file P1 (again, the example data are fictitious):

Service Year	Person Key	Coverage Class	CCHG Cat.	CCHG Cat. Desc.	Utilization Type	Total Allowed	Total Paid	Claim Line Count
2016	999999	MED	106	HIV	Nursing Facility	$13,579	$10,864	3
2016	999999	MED	106	HIV	Clinic/Office	$ 500	$ -	2
2016	999999	MED	106	HIV	Hospital Outpatient	$ 1,525	$ 875	3

The adversary has matched carrier, product type, and member demographic characteristics to all records of this member—a member of a rare and sensitive diagnostic category—and the records of at least ten other individuals in the file, greatly increasing the amount of indirectly identifying information known about the members in the matched group.

However, if only a 50% sample of members appear in the person-level file, the maximum probability that file P1 includes all members of any group of 11 or more members whose data are summarized on a single row of file A2 is

$$1/(2^{11}) = 1/2048$$

(because there is a 50%, or 1-out-of-2, chance any individual member is in the sample data set), rendering such matches unlikely. In addition, the adversary cannot assume any such unique match found in a random sample data release to be unique in the population data set, and therefore cannot necessarily associate the additional information with the identified person-level record set.

BerryDunn also considered the potential risks of linking the person-level enhanced public use files to the current claim-level Safe Harbor public use files. While in theory an adversary finding a unique exact match on quantitative variables between a group of claims on the claim-level public use files and the proposed person-level file would have effectively attached a member identifier to that group of claims, the computing power and programming skills to develop and execute such an attack present a significant barrier. Moreover, while such a match would provide access to a number of additional quasi-identifiers (e.g., specific year of age, gender, specific procedure codes, secondary diagnosis codes, in some cases some limited provider information), no new knowledge would be gained besides the fact that a given set of claims belonged to one individual. This is because all information other than the encrypted member identifier on the person-level enhanced public use file is available on, or derivable from, data present on the claim-level public use files. Given these considerations, and taking into account the level of *k*-anonymization applied in the enhanced public use file, this potential linkage does not present a significant additional risk of re-identification.

It is clear that some recipients of NH CHIS limited use data sets could make successful linkages to the enhanced public use files. However, given that all information available on the enhanced public use files is available in the Consolidated NH CHIS, and that such linkage would violate users' data use agreements, such linkage is not considered a significant re-identification risk. Similarly, individuals with access to other sources of health care data, such as providers or health insurers, are bound by HIPAA against data linkage and re-identification attempts.

Finally, a de-identification expert must also consider the risk of linkage to outside data sources, such as voter registration or vital records databases. In the present case, the lack of specific demographic data and the application of *k*-anonymization in the person-level enhanced public use files, as well as the fact that the person-level enhanced public use files are a 50% sample of a sample (the commercially insured population covered by the subset of carriers meeting New Hampshire's data submission requirements) lead to the conclusion that the risk of re-identification based on linkage to such outside sources is very—and sufficiently—small.

Sampling reduces minimum equivalence class requirements

When a sample of personal data are taken prior to de-identification, the equivalence class size required for de-identification can be reduced in proportion to the size of the sample.[31] That is, the required equivalence class size for the NH CHIS person-level expense public use files can be reduced by half to a standard of $k \geq 10$. However, the current design uses generalization and

suppression to achieve a minimum equivalence class size of $k \geq 11$; this slightly more conservative rule standardizes the aggregated file cell size and person-level file equivalence size requirements in the design.

Truncation of Extreme Data Values

Although quantitative data elements such as allowed charges, paid dollars, claim counts, drug days supply, etc., are neither direct identifiers nor quasi-identifiers, extreme values reported in these fields are potential additional sources of uniqueness or near-uniqueness, and in the most extreme cases may pose a significant re-identification risk. For example, a single case of a new, very expensive treatment being tried for a certain medical condition could make the news, or a potential adversary could be aware that a neighbor with a particular chronic condition had an extremely large number of encounters with the health care system over the past year.

To mitigate these risks, BerryDunn capped total allowed medical charges and paid claim amounts at $250,000 at the member level, and claim line counts at 500. The $250,000 figure was chosen because it is a common actuarial threshold for identifying extreme high-cost claimants. Spending for approximately 370 members in the medical population test file exceeds the threshold. The claim line threshold was set such that a similar number of members was affected.

After summarizing expenses and claim line counts at the member level, the following data truncation algorithm was applied for all members exceeding the threshold amount for any of the three quantitative variables.

If a user is the *only* user in the 50% sample over one of these caps in a Chronic Condition Hierarchical Grouping, that member's data are suppressed (i.e., not reported in the person-level medical expense public use enhancement files). If there are multiple members in the diagnostic grouping over the cap, their spending and/or claims are capped at the above limits at the member level, and the limit amount is distributed over the utilization categories in proportion to the actual spending across the utilization categories before truncation.

For example, if a member has $1 million in allowed charges, $750,000 of which is Hospital Inpatient utilization, $200,000 of which is Hospital Outpatient, and $50,000 of which is Clinic Office, then the reported spending would be $250,000*(.75) = $187,500 Hospital Inpatient, $250,000*(.2) = $50,000 Hospital Outpatient, and $250,000*(.05) = $12,5000 Clinic Office. A companion report should be produced showing the amounts of suppressed and capped dollar amounts. In testing, data for members accounting for $3,189,243 in allowed charges and $4,531,703 paid dollars were suppressed because a single member exceeded the cap in a diagnostic grouping.

Allowed charges of $21,808,770 and $22,322,196 in paid dollars were suppressed for being over the cap for members whose expenses were reported in the test file at the $250,000 cap.

Similar summaries of suppressed claim line counts and users should *not* be reported. If the totals for these fields do not reach the reportability threshold, such a summary would fail the aggregated

data reportability standards. If, on the other hand, the totals are reported where they do meet the threshold, their absence in the companion documentation to other files (where the threshold was not met) would be tantamount to reporting totals beneath the threshold.

A floor of zero dollars and zero claim lines is also applied at the member level. All utilization category amounts for a member with a zero or negative total are set to zero in these cases. The presence of zero or negative values is not considered potentially identifying *unless* a member is the only member in a Chronic Condition Hierarchical Grouping with total allowed *and* paid expenses less than or equal to zero in the sample. The possible presence of exactly one person in a diagnostic grouping with expenses summing to a carrier refund is considered potentially identifying, and the member's data are suppressed. This condition resulted in an immaterial amount of suppression in testing.

Please note that, because data truncation is performed at the member level, utilization category-level records with negative amounts can and do appear for members with positive overall totals in a given quantitative field. That is, if a member has -$234 allowed reported for utilization category "Clinic/Office" and $567 allowed reported for "Hospital Outpatient" and no other utilization, neither record is affected by the allowed amount truncation rules.

Suppression for violating the cap or floor is performed before *k*-anonymization; if, after *k*-anonymization, the generalization and suppression algorithms have created additional cases of a single member violating a cap or floor within a Chronic Condition Hierarchical Grouping, this is not considered potentially identifying and the member's data are *not* suppressed because, by definition, there were additional members in the diagnostic grouping over the cap prior to *k*-anonymization (and still more such cases are likely present in the other half of the population data). This rule prevents a potentially infinite recursion of the suppression and generalization process.

***K*-anonymization (Generalization Algorithm)**

The first step in *k*-anonymization is to run a query summing allowed charges, paid amounts, and claim line counts by member (PERSON_KEY), Chronic Condition Hierarchical Grouping (CCHG_CAT), and utilization type (UTILIZATION_TYPE) for all medical claims in the test data sample and test *k*-anonymity at the Chronic Condition Hierarchical Grouping and utilization type level (or "row") level. Any Chronic Condition Hierarchical Grouping and utilization type combination must occur for at least 11 PERSON_KEYs to be reported without generalization.

In testing, there were $192,000 (0.02% of test data volume after random sampling and data truncation) in allowed charges volume for diagnostic grouping/utilization category combinations with less than 11 members. BerryDunn applied the following generalization algorithm to address these cases. After each of the numbered generalizations, the data were re-summarized by the "group-by" variables and the results re-checked for adequate equivalence class size:

1. Utilization Type: UTILIZATION_TYPE was recoded as "OUM = Other, Unknown, or Masked" for rows with inadequate equivalence class sizes. This generalization resolved equivalence class size for records accounting for 80% ($153,000 allowed) of the unreportable volume, leaving an immaterial $39,000 that required further generalization.
2. Chronic Condition Hierarchical Grouping: CCHG_Cat was recoded as "999 = Unknown or Masked CCHG" for all rows for members with any rows with remaining inadequate equivalence class sizes (this is required to maintain the unique assignment of diagnostic groupings to members). This generalization resolved all row-level equivalence classes.

Because the diagnostic category and, in particular, utilization category groupings are so broad, with only eight service types, including "Other, Unknown, or Masked," utilization patterns are not considered potentially identifying in the vast majority of cases, and achieving *k*-anonymity at the "row" level is considered sufficient.

However, BerryDunn defined the following subset of member records that require an additional *k*-anonymization pass at the Chronic Condition Hierarchical Grouping category, utilization type, and utilization pattern level. That is, a PERSON_KEY's exact set of rows (e.g., a Healthy Male (65+) with Clinic/Office, Ambulatory Surgery, and Hospital Outpatient utilization rows) must be found for at least 10 other PERSON_KEYs to be reported without further generalization when *both* of the following are true:

1. The user's Chronic Condition Hierarchical Grouping includes less than 1% of the service user population
2. The user has less than four utilization types. "Sparse" utilization patterns are considered more potentially identifying because they are more likely to result in unusual patterns that have a higher probability of uniqueness and that a potential adversary might realistically know (e.g., an individual with only Home Health/Hospice services).

Using this standard, $4.3 million (0.4% of test data volume after random sampling and data truncation) allowed in the test data fall in utilization pattern equivalence classes of $k < 11$ before generalization. For these cases, generalization proceeded as follows:

1. Utilization Type: UTILIZATION_TYPE was recoded as "OUM = Other, Unknown, or Masked" for members with inadequate utilization pattern equivalence class sizes. This generalization resolves utilization pattern equivalence class size in all cases.
2. Chronic Condition Hierarchical Grouping: CCHG_Cat would be recoded as "999 = Unknown or Masked CCHG" for any remaining cases of inadequate utilization pattern equivalence class size.

Please note that it is unlikely that the utilization type generalization pass will not resolve all utilization pattern equivalence classes, as it sets utilization category equal to "Other, Unknown, or

Masked" for all rows for the affected members, such that all members within a diagnostic grouping requiring utilization pattern equivalence class generalization will be included in the same (common) utilization pattern equivalence class after the utilization category generalization pass.

Data for any members in the utilization pattern equivalence class testing population whose utilization pattern equivalence classes are not resolved must be suppressed from the production report, and their corresponding expense dollars (but not claim line or user counts) reported in a companion report.

A third and final generalization pass may be required if utilization pattern generalization reduces the number of members with any given Chronic Condition Hierarchical Grouping/utilization category combination to less than 11 (that is, if utilization pattern generalization causes some records to "lose" row-level equivalence class adequacy). For these rows, utilization category should be generalized, followed by diagnostic grouping for any members remaining with rows with inadequate equivalence class sizes. Data must be suppressed for any members with any rows with inadequate equivalence class size after this final generalization pass.

Rows comprising $40,000 allowed "lost" row-level equivalence class adequacy after utilization pattern generalization. Generalizing utilization category for these rows resolved all such cases.

The final test file includes approximately 660,000 rows for approximately 281,000 unique PERSON_KEYs.

BerryDunn has provided its programming code to DHHS, NHID, and the departments' data vendor to provide further detailed documentation of this methodology and facilitate initial production of the enhanced public use files.

Table 9
Population Cost by Member – Pharmacy (File P2)

Row Definition Fields	Order of Generalization	Field Name	Common Name	NH CHIS Data Element ID	Type	Length	Description	On Current Public Use Files?
✓		service_year	Service Year	MC059, PC032	Number	4		Yes
✓		person_key	Hashed person key	N/A	TBD	TBD	Must be "re-hashed" for every release to avoid intertemporal linkage	No
		coverage_class	Coverage Class	MC899	String	3	This field indicated the type of record. For all medical claims records, this value will be MED. Pharmacy Claims are PHM.	Yes
✓	1	CCHG_Cat	Chronic Condition Hierarchical Grouping Code	N/A	String	4	Chronic Condition Hierarchical Groups organize medical utilization and costs in a clinically relevant manner. Milliman's CCHGs assign patients to unique categories using a clinically relevant hierarchy that groups similar patients in the same group based on how doctors make treatment decisions. i.e. 101 - Major Psychosis, 114 - COPD, or 130 - Healthy Male (41-64). The population cost by member file reflects the following changes to Milliman's CCHG_Cat assignment: -The ten "Healthy Male" and "Healthy Female" age group categories for ages 65 and over have been recoded to the categories "145...Healthy Male (65+)" and "146...Healthy Female (65+)" -All claims for a member in the year are assigned the member's most recent valid CCHG_Cat (as defined by claim line service from date), if any. -Null CCHG_Cat values and CCHG_Cat values masked to achieve adequate equivalence class size (*k*>=20) are set to "999...Unknown or Masked CCHG"	Yes
		CCHG_Cat_Desc	Chronic Condition Hierarchical Grouping Description	N/A	String	92	Description of the CCHG code (e.g., "Major Psychosis," "Healthy Female (6-15)").	Yes
		total_allowed	Total Allowed	MC063, MC067, MC066, MC065, PC036, PC042, PC041, PC040	Decimal	10,2	This field identifies the sum of all plan and member payments for services rendered to the individual in the reporting year. This is a money field containing dollars and cents.	Not Derivable at Person Level
		total_paid	Total Paid	MC063, PC036, MC064	Decimal	10,2	This field identifies the sum of all plan payments for services rendered to the individual in the reporting year. This is a money field containing dollars and cents.	Not Derivable at Person Level
		total_days_supply	Total days' supply	PC034	Decimal	10,2	This field identifies the total days' supply of paid utilization. No absolute value < 11 can be reported in this field.	Not Derivable at Person Level
		claim_line_count	Count of Claim Lines Summarized per Row	N/A	Number	10	Number of paid claim line items summarized in the present row. Paid claim line items (SV_STAT = "P") are summed with a value of 1. Reversals (SV_STAT = "R") are summed with a value of -1.	Not Derivable at Person Level

5.2 Population Pharmacy Cost by Member File (File P2)

Table 9 above shows the proposed layout of the population pharmacy cost by member file (File P2). The file shows commercial allowed charges, carrier paid amounts, drug days supply, and claim line counts summarized by individual [xix] and general diagnostic category. The presence of an encrypted person-level identifier (the PERSON_KEY field) in the file renders the content personal information, and re-identification risks must be managed as such.

The academic literature on the de-identification of personal health information suggests a maximum re-identifiability risk of 5% for a public data set's most vulnerable record.[32]

As noted above, this standard requires removal or encryption of any direct identifiers and the creation of adequate equivalence classes. A simple metric for assessing re-identification risk, or the probability of re-identification (Pr(ID)), posits that re-identification risk is no greater than the inverse of the size of the minimum equivalence class; that is, given that a probability of re-identification less than or equal to 5% is equivalent to:

$$\Pr(\mathrm{ID}) \leq 1/20$$

requiring equivalence classes of 20 for all combinations of quasi-identifiers in a population data set and effectively encrypting direct identifiers meets a generally accepted standard of de-identification of person-level health care data for public use.

Testing the sizes of equivalence classes is often referred to in the literature as "testing *k*-anonymity," where *k* is the number of members required to form an adequately de-identified equivalence class.

The proposed design for file P2 includes the quasi-identifier Chronic Condition Hierarchical Grouping Category (CCHG_CAT) and its description. These categories are value-added data developed by DHHS's NH CHIS data vendor that "organize medical utilization and costs in a clinically relevant manner. [The] CCHGs assign patients to unique categories using a clinically relevant hierarchy that groups similar patients in the same group based on how doctors make treatment decisions. i.e. 101 – Major Psychosis, 114 – COPD, or 130 – Healthy Male (41 – 64)."[33]

In the test file, 70% of PERSON_KEYs had at least one claim assigned to a non-null Chronic Condition Hierarchical Grouping. For those PERSON_KEYs with multiple non-null Chronic Condition Hierarchical Groupings reported, BerryDunn assigned a single, unique Chronic Condition Hierarchical Grouping to

[xix] In the report designs, a distinct PERSON_KEY defines a unique individual. Throughout this discussion, the terms "PERSON_KEY," "member," and "individual" are used interchangeably. The Consolidated NH CHIS documentation defines this field as "the key that identifies a unique person within the data warehouse." If a single value of PERSON_KEY appears under multiple payer codes, the volume for all payers associated with the PERSON_KEY value are summed together on the lines reporting that PERSON_KEY. On the other hand, if a person's claims or eligibility are reported under multiple PERSON_KEYs in the NH CHIS medical data, that person's data will be reported under multiple PERSON_KEYs in the population cost file, as if the records belonged to multiple people.

all claims for a single PERSON_KEY by assigning all claims to the last valid Chronic Condition Hierarchical Grouping reported in the year by service date.[xx]

File P2 therefore includes one and only one record for each reported member, summarizing all reported pharmacy utilization on that row.

The following sub-sections detail the risk-management steps required to produce the de-identified medical expense by member file.

Re-encryption of PERSON_KEY

As noted above, file P2 includes the Consolidated NH CHIS member identifier PERSON_KEY. Although PERSON_KEY, as it appears in the Consolidated NH CHIS, is already encrypted and is therefore not a direct member identifier, the enhanced public use file methodology requires the following re-encryption to minimize the risk of member re-identification by linking between NH CHIS data sources:

1. All values of PERSON_KEY must be randomly assigned a new value (rather than perturbing each value by a simple transformation, or "hash"). DHHS's data vendor should securely maintain a mapping of the Consolidated NH CHIS PERSON_KEY values to the population cost by member files' values.
2. PERSON_KEY values for the medical and pharmacy versions of the population cost by member files should be reassigned separately, such that an individual's data may *not* be linked across the two files.
3. Each year's file should be encrypted separately such that an individual's data may *not* be linked across files for multiple years.

Numbers 2 and 3 above are required because *k*-anonymity is tested in isolation for the data in each coverage class (medical or pharmacy) and each year. That is, the methodology does not require that the data reported in the 2016 pharmacy file for a given individual form adequate equivalence classes when combined with data for that individual in the 2016 medical file, or in the 2017 pharmacy file.

Random Sampling

The first step in producing the file is to identify a 50% random sample of members with paid utilization and limit the analysis data set to claim lines associated with only the members in the sample. This choice has two main advantages for the present purpose. First, sampling provides a measure of security against attempts to re-identify observations in the person-level files by linking to additional data in other data sources. Second, when a sample of personal data are taken prior to de-identification,

[xx]If a PERSON_KEY had a single claim each in January, September, and December 2016, if the January 2016 claim reported a CCHG_CAT of "125 = Other chronic conditions," the September 2016 claim reported "130 = Healthy Male (41-64), and the December claim reported a null CCHG_CAT, CCHG_CAT "130 = Healthy Male (41-64) would be assigned to all three claims in this report.

the equivalence class size required for de-identification can be reduced in proportion to the size of the sample.[34]

After the 50% random sample data set is created, total expenses, claim lines, and users in the sample file should be checked against the population totals to ascertain that the random sampling process has yielded a data set including half the volume of the population file on those aggregated measures. BerryDunn recommends the sample be tested against a standard of including 50% of total allowed charges, payments, claim counts, drug days supply, and distinct users when the ratio of the totals in the sample to the totals in the population file are rounded to the 1% (hundredth) place.

BerryDunn's resulting pharmacy claims 50% sample test file included:

- 4,166,697 claim lines, less reversals (50% of 8,331,465 claim lines in the full test file)
- $507,232,275 allowed (50% of $1,017,093,185)
- $469,794,205 paid (50% of $941,396,349)
- 305,778 distinct users (50% of 611,611)

The following sub-sections describe the advantageous consequences of sampling in detail.

Sampling protects against file linkage

Sampling provides a measure of security against attempts to re-identify observations in the person-level files by linking to additional data in other data sources.

BerryDunn identified no significant linkage risks between file P2 and other files in the proposed enhanced public use file design because Chronic Condition Hierarchical Grouping does not appear in any other pharmacy files in the design.

BerryDunn also considered the potential risks of linking the person-level enhanced public use files to the current claim-level Safe Harbor public use files. While in theory an adversary finding a unique exact match on quantitative variables between a group of claims on the claim-level public use files and the proposed person-level file would have effectively attached a member identifier to that group of claims, the computing power and programming skills to develop and execute such an attack present a significant barrier. Moreover, while such a match would provide access to a number of additional quasi-identifiers such as specific year of age, gender, drug code and name, no new knowledge besides the fact that a given set of claims belonged to one individual would be gained. This is because all information other than the encrypted member identifier on the person-level enhanced public use file is available on, or derivable from, data present on the claim-level public use files. Given these considerations, and taking into account the level of *k*-anonymization applied in the enhanced public use file, this potential linkage presents no significant additional risk of re-identification.

It is clear that some recipients of NH CHIS limited use data sets could make successful linkages to the enhanced public use files. However, given that all information available on the enhanced public use files is available in the Consolidated NH CHIS, and that such linkage would violate users' data use agreements, such linkage is not considered a significant re-identification risk. Similarly, individuals

with access to other sources of health care data, such as providers or health insurers, are bound by HIPAA against data linkage and re-identification attempts.

Finally, a de-identification expert must also consider the risk of linkage to outside data sources, such as voter registration or vital records databases. In the present case, the lack of specific demographic data and the application of *k*-anonymization in the person-level enhanced public use files, and the fact that the person-level enhanced public use files are a 50% sample of a sample (the commercially-insured population covered by the subset of carriers meeting New Hampshire's data submission requirements) lead to the conclusion that the risk of re-identification based on linkage to such outside sources is very—and sufficiently—small.

Sampling reduces minimum equivalence class requirements

When a sample of personal data are taken prior to de-identification, the equivalence class size required for de-identification can be reduced in proportion to the size of the sample.[35] That is, the required equivalence class size for the NH CHIS person-level expense public use files can be reduced by half to a standard of $k \geq 10$. However, the current design uses generalization and suppression to achieve a minimum equivalence class size of $k \geq 11$; this slightly more conservative rule standardizes the aggregated file cell size and person-level file equivalence size requirements in the design.

Truncation of Extreme Data Values

Although quantitative data elements such as allowed charges, paid dollars, claim counts, drug days supply, etc., are neither direct identifiers nor quasi-identifiers, extreme values reported in these fields are potential additional sources of uniqueness or near-uniqueness, and in the most extreme cases may pose a significant re-identification risk. For example, a single case of a new, very expensive drug being tried for a certain condition could make the news, or a potential adversary could be aware that a neighbor with a particular chronic condition had an extremely large number of prescriptions filled over the past year.

To mitigate these risks, BerryDunn capped total allowed pharmacy charges and paid claim amounts at $125,000 at the member level, drug days supply at 7500, and claim line counts at 175. The numbers $125,000, 7500, and 175 were chosen because these thresholds affected a similar number of pharmacy members as were affected by the higher medical thresholds. Approximately 390 members in the pharmacy population test file exceed a given threshold.

After summarizing pharmacy expenses, drug days supply, and claim line counts at the member level, the following data truncation algorithm was applied for all members exceeding the threshold amount for any of the four quantitative variables.

If a user is the *only* user in the 50% sample over one of these caps in a Chronic Condition Hierarchical Grouping, that member's data are suppressed (i.e., not reported in file P2). If there are multiple members in the diagnostic grouping over the cap, their spending and/or claims are capped at the above limits. In testing, data for members accounting for $1,695,011 in allowed charges and $1,736,031 paid dollars were suppressed because a single member exceeded the cap in a diagnostic grouping.

Allowed charges of $14,942,548 and $14,774,110 in paid dollars were suppressed for being over the cap for members whose expenses were reported in the test file at the $125,000 cap.

Similar summaries of suppressed drug days supply, claim line counts, and users should *not* be reported. If the totals for these fields do not reach the reportability threshold, such a summary would fail the aggregated data reportability standards. If, on the other hand, the totals are reported where they do meet the threshold, their absence in the companion documentation to other files (where the threshold was not met) would be tantamount to reporting totals beneath the threshold.

A floor of zero dollars, zero days supply, and zero claim lines is also applied at the member level. All amounts for a member with a zero or negative total are set to zero in these cases. The presence of zero or negative values is not considered potentially identifying *unless* a member is the only member in a Chronic Condition Hierarchical Grouping with total allowed *and* paid expenses less than or equal to zero in the sample. The possible presence of exactly one person in a diagnostic grouping with expenses summing to a carrier refund is considered potentially identifying, and the member's data are suppressed. This condition resulted in an immaterial amount of suppression in testing.

Suppression for violating the cap or floor is performed before *k*-anonymization. If, after *k*-anonymization, the generalization and suppression algorithms have created additional cases of a single member violating a cap or floor within a Chronic Condition Hierarchical Grouping, this is not considered potentially identifying and the member's data are *not* suppressed because, by definition, there were additional members in the diagnostic grouping over the cap prior to *k*-anonymization (and still more such cases are likely present in the other half of the data). This rule prevents a potentially infinite recursion of the suppression and generalization process.

K-anonymization (Generalization Algorithm)

The first step in *k*-anonymization is to run a query summing allowed charges, paid amounts, drug days supply, and claim line counts by member (PERSON_KEY) and Chronic Condition Hierarchical Grouping (CCHG_CAT) for all pharmacy claims in the test data sample and test *k*-anonymity at the Chronic Condition Hierarchical Grouping level. Any Chronic Condition Hierarchical Grouping must occur for at least 11 PERSON_KEYs to be reported without generalization.

In testing, all records met the equivalence class threshold without generalization. Should any Chronic Condition Hierarchical Groupings fail equivalence class adequacy in the future, records for that diagnostic category should be recoded "999 = Unknown or Masked CCHG." As long as a significant percentage of members in the pharmacy data are not associated with a valid Chronic Condition Hierarchical Grouping, this generalization is expected to resolve all equivalence classes.

The final test file includes 305,761 rows for 305,761 unique PERSON_KEYs.

6 Looking forward

6.1 Recommended frequency of review

Results of each annual production run must be monitored for continued compliance with the de-identification standards described above, and the results assessed for continued analytical utility.

In particular, the results should be reviewed each year for significant changes (increases) in the volume of generalized, capped, and most importantly, suppressed data generated. The following sub-sections describe likely causes for such changes and recommendations for maintaining compliance and utility. The programming code provided to the departments and their data vendor outputs statistics summarizing the effect of each step in the de-identification process that can be reviewed for changes against prior runs.

6.2 Foreseeable future challenges

The most likely future challenge to the design is the potential evolution of NH CHIS data in ways that decrease the size of the data universe, or specific subsets of the data universe, such as:

- Any ongoing effects of the March 2016 *Gobeille v. Liberty Mutual Ins. Co.* United States Supreme Court decision resulting in further deterioration of the self-insured data
- Possible future instability in the individual market arising from legislative changes
- Submission errors on the part of data submitters

Reductions in the size of the input data universe will necessarily increase the difficulty of achieving any given cell size or equivalence class standard.

Changes to the HIPAA Privacy Rule or state laws may also require changes to the design. Given that DHHS and NHID already closely monitor the relevant regulatory landscape, the departments are well-positioned to identify and react to any such changes in a timely manner.

It is also possible that technological changes, analytical breakthroughs, or high-profile privacy breaches could result in changes to commonly-accepted de-identification standards. BerryDunn recommends the departments (or a contractor) conduct a review of new de-identification literature at least every 36 months and on an *ad hoc* basis in the wake of relevant news events or regulatory changes.

6.3 Recommended techniques and approaches for continued compliance and analytical utility

Correct implementation of the recommended methodology should result in enhanced public use files meeting the de-identification standards described above. In the event of changes to commonly-accepted de-identification standards, the design will require review by an expert to assess its continued compliance with current standards and develop solutions to mitigate any emerging deficiencies. The public use files should be made unavailable to new users pending the review.

In the more likely case that the evolution of NH CHIS submissions results in significant increases in generalization or suppression, BerryDunn recommends the following steps:

- Identify the step or steps in the de-identification process yielding deteriorating results
- Identify the record set affected by the change in the input data
- Assess what data elements and which domain values are driving problematic results
- In the case of increased suppression due to equivalence classes below the minimum:
    - Consider possible recodes of field domain values to create adequate equivalence classes
    - If recodes do not mitigate the problem, consider removing a quasi-identifier(s) from the design. While it may seem intuitive to drop the quasi-identifier(s) that appear to be driving the increase in suppression, the analyst should also consider analytical utility in developing an order for removing variables. If a high-analytical-value field is the driver of the deteriorating results, removing a quasi-identifier of lesser analytical utility may ameliorate the problem.
- In the case of increased suppression of data owing to data truncation, the analyst should consider whether trends in health care utilization, spending, and inflation suggest a reassessment of the claim and dollar caps.

For example, consider the possibility of a significant contraction of the New Hampshire individual health insurance market resulting in significant increases in generalization or suppression in the medical expense by product type, funding source, and market segment file (file A2). The analyst might recode the market segment (MKT_SEG) quasi-identifier to include individual market products with small group or the "other, unknown, or masked" segment, especially if the individual market was projected to continue deteriorating in coming years.

If such a recode is not advisable, or does not solve the problem, the analyst should then consider whether analytical utility is better served by publishing the remaining reportable data "as-is" or in assessing quasi-identifiers to be removed from the design. For example, the analyst may decide that, given the large size of the age groups (0 – 25 years, 26 – 64 years, and 65+ years) in the age group (AGE_GROUP) field, the reduction in generalization or suppression in other fields resulting from removing age group from the design presents a worthwhile trade-off in analytical utility.

7 Conclusion

Since 2012, DHHS has provided NH CHIS public use data sets of medical, pharmacy, and dental claims de-identified utilizing the HIPAA Privacy Rule Safe Harbor method to members of the public upon written request. Seeking to improve the utility of these data sets for secondary purposes, NHID and DHHS engaged BerryDunn to provide recommendations for the enhancement of the NH CHIS public use data sets utilizing the HIPAA Privacy Rule Expert Determination method, or statistical method.

In assessing the trade-offs between re-identification risk and analytical utility, BerryDunn concluded that the amount of data generalization and suppression required to enhance claim-line level release data for public use beyond the strictures of the Safe Harbor method while maintaining adequate de-identification under the Expert Determination method resulted in data that were not analytically useful. As an alternative approach, BerryDunn developed the suite of 11 files described above to be released alongside the current claim-level Safe Harbor public use data sets.

This approach maintains adequate de-identification while providing several files highly useful for answering certain common questions in health care analysis. In addition, the new files can be stored and manipulated in most common spreadsheet applications, rather than requiring that the user have statistical programming skills in order to analyze the data.

BerryDunn designed the proposed enhanced NH CHIS public use files and production process to aid NHID and DHHS as they build upon New Hampshire's tradition of health care data transparency. With this proposed system, the departments will improve the utility and accessibility of their public health care analytical resources and meet established de-identification standards for a reasonable expenditure of resources and ongoing effort.

Endnotes

[1] United States Department of Health and Human Services Office for Civil Rights. *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.* November 26, 2012. Accessed 18 September 2017: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf.

[2] NH Title XXXVII, Chapter 420-G, Portability, Availability, and Renewability of Health Coverage. Section 420-G:11-a.

[3] *Op. cit.* United States Department of Health and Human Services Office for Civil Rights. *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.*

[4] United States Department of Health and Human Services Office for Civil Rights. *Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.* November 26, 2012. Accessed 18 September 2017: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf.

[5] *Ibid.*

[6] *Ibid.*

[7] *Ibid.*

[8] *Ibid.*

[9] *Ibid.*

[10] *Ibid.*

[11] El Emam, K. *Guide to the De-Identification of Personal Health Information.* Boca Raton, FL. CRC Press, 2013. p.159.

[12] El Emam K, Arbuckle L. *Anonymizing Health Data: Case Studies and Methods to Get you Started.* O'Reilly Media, Inc., Sebastopol, CA 2013; p. 28.

[13] *Op. cit.* El Emam K, Arbuckle L. *Anonymizing Health Data: Case Studies and Methods to Get you Started.* p. 28.

[14] *Op. cit.* El Emam, K. *Guide to the De-Identification of Personal Health Information.* p. 159.

[15] El Emam K, Arbuckle L, Koru G, Eze B, Gaudette L, Neri E, Ross S, Howard J, Gluck J. *De-identification Methods for Open Health Data: The Case of the Heritage Health Prize Claims Dataset.* J Med Internet Res 2012:14(1):e33. Accessed 20 September 2017: http://www.jmir.org/2012/1/e33/.

[16] *Ibid.*

[17] *Op. cit.* El Emam K, Arbuckle L. *Anonymizing Health Data: Case Studies and Methods to Get you Started.* P. 29.

[18] *Op. cit.* El Emam, K. *Guide to the De-Identification of Personal Health Information.* pp. 164-5.

[19] *Ibid,* pp. 279-80.

[20] Research Data Assistance Center. *CMS Cell Size Suppression Policy.* Current Version Date: 05/08/2017. Accessed 18 September 2017: https://www.resdac.org/resconnect/articles/26.

[21] *Op. cit.* El Emam K, Arbuckle L, Koru G, Eze B, Gaudette L, Neri E, Ross S, Howard J, Gluck J. *De-identification Methods for Open Health Data: The Case of the Heritage Health Prize Claims Dataset.*

[22] *Op. cit.* El Emam, K. *Guide to the De-Identification of Personal Health Information.* p.132.

[23] *Ibid.* p. 128.

[24] *Ibid.* p. 128-129.

[25] *Op. cit.* El Emam K, Arbuckle L, Koru G, Eze B, Gaudette L, Neri E, Ross S, Howard J, Gluck J. *De-identification Methods for Open Health Data: The Case of the Heritage Health Prize Claims Dataset.*

[26] Research Data Assistance Center. *CMS Cell Size Suppression Policy.* Current Version Date: 05/08/2017. Accessed 18 September 2017: https://www.resdac.org/resconnect/articles/26.

[27] *Op. cit.* El Emam K, Arbuckle L, Koru G, Eze B, Gaudette L, Neri E, Ross S, Howard J, Gluck J. *De-identification Methods for Open Health Data: The Case of the Heritage Health Prize Claims Dataset.*

[28] *Op. cit.* El Emam, K. *Guide to the De-Identification of Personal Health Information.* p.225-226.

[29] New Hampshire Comprehensive Health Care Information System. *The New Hampshire Comprehensive Health Care Information System (NH CHIS) Consolidated Data Dictionary, Version 3.* April 2017, p. 40.

[30] *Op. cit.* El Emam K, Arbuckle L, Koru G, Eze B, Gaudette L, Neri E, Ross S, Howard J, Gluck J. *De-identification Methods for Open Health Data: The Case of the Heritage Health Prize Claims Dataset.*

[31] *Ibid.*

[32] *Op. cit.* El Emam, K. *Guide to the De-Identification of Personal Health Information.* p.225-226.

[33] *Op. cit.* New Hampshire Comprehensive Health Care Information System. *The New Hampshire Comprehensive Health Care Information System (NH CHIS) Consolidated Data Dictionary, Version 3.* April 2017, p. 40.

[34] *Op. cit.* El Emam K, Arbuckle L, Koru G, Eze B, Gaudette L, Neri E, Ross S, Howard J, Gluck J. *De-identification Methods for Open Health Data: The Case of the Heritage Health Prize Claims Dataset.*

[35] *Ibid.*