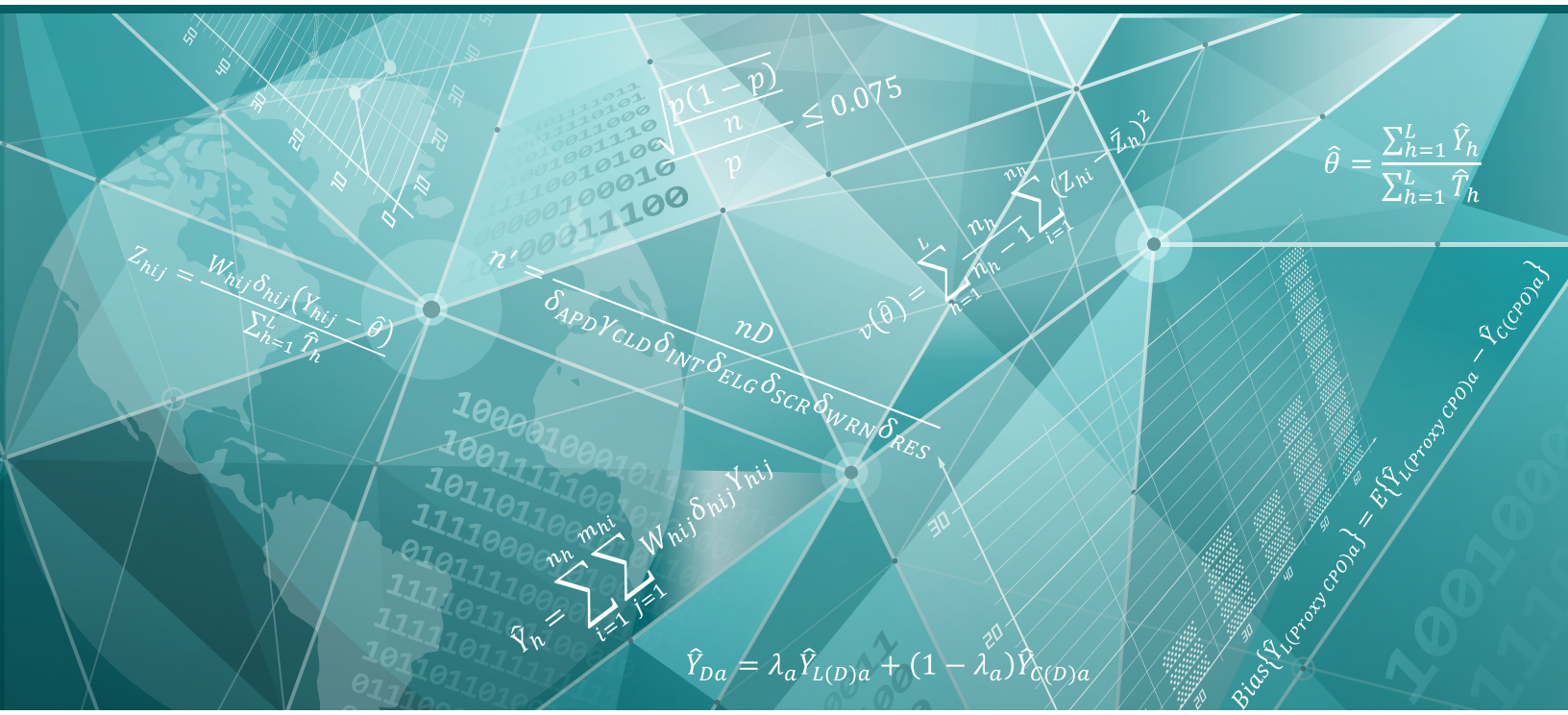
NATIONAL CENTER FOR HEALTH STATISTICS

Vital and Health Statistics

Series 2, Number 189

October 2021



Machine Learning for Medical Coding in Healthcare Surveys

Data Evaluation and Methods Research

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Centers for Disease Control and Prevention
National Center for Health Statistics

NCHS reports can be downloaded from: https://www.cdc.gov/nchs/products/index.htm.

Copyright information

All material appearing in this report is in the public domain and may be reproduced or copied without permission; citation as to source, however, is appreciated.

Suggested citation

Lucas CA, Hadley E, Chew R, Nance J, Baumgartner P, Thissen MR, et al. Machine learning for medical coding in healthcare surveys. National Center for Health Statistics. Vital Health Stat 2(189). 2021. DOI: https://dx.doi.org/10.15620/cdc:109828.

For sale by the U.S. Government Publishing Office
Superintendent of Documents
Mail Stop: SSOP
Washington, DC 20401–0001
Printed on acid-free paper.

NATIONAL CENTER FOR HEALTH STATISTICS

Vital and Health Statistics

Series 2, Number 189 October 2021

Machine Learning for Medical Coding in Healthcare Surveys

Data Evaluation and Methods Research

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Centers for Disease Control and Prevention
National Center for Health Statistics

Hyattsville, Maryland
October 2021

National Center for Health Statistics

Brian C. Moyer, Ph.D., *Director*

Amy M. Branum, Ph.D., *Associate Director for Science*

Division of Health Care Statistics

Carol J. DeFrances, Ph.D., *Acting Director*

Alexander Strashny, Ph.D., *Associate Director for Science*

Contents

Text Figures

Detailed Tables

Contents—Con.

Acknowledgments

The authors would like to thank HealthCare Resolution Services for providing context on the medical coding process, and Research Triangle Institute systems staff for their many contributions to this work.

The authors also gratefully thank Amy Blum and Kellina Phan of the National Center for Health Statistics (NCHS) for their tireless review of the verbatim data referenced in this report and their expertise in medical coding, and Susan Schappert, also of NCHS, for her guidance and effort in readying the verbatim data for review.

Machine Learning for Medical Coding in Healthcare Surveys

by Christine A. Lucas, Ph.D., M.P.H., M.S.W., National Center for Health Statistics; Emily Hadley, M.S., Robert Chew, M.S., Jason Nance, M.C.S., Peter Baumgartner, M.S., M. Rita Thissen, M.S., David M. Plotner, M.S., Christine Carr, M.A., RTI International; and Aerian Tatum, D.B.A., M.S., RHIA, CCS, HealthCare Resolution Services

Abstract

Objective

Medical coding, or the translation of healthcare information into numeric codes, is expensive and time intensive. This exploratory study evaluates the use of machine learning classifiers to perform automated medical coding for large statistical healthcare surveys.

Methods

This research used medically coded data from the Emergency Department portion of the 2016 and 2017 National Hospital Ambulatory Medical Care Survey (NHAMCS-ED). Natural language processing classifiers were developed to assign medical codes using verbatim text from patient visits as inputs. Medical codes assigned included three-digit truncated 10th Revision of the *International Statistical Classification of Diseases and Related Health Problems, Clinical Modification* (ICD-10-CM) codes for diagnoses (DIAG) and cause of injury (CAUSE), as well as the full length NCHS reason for visit (RFV) classification codes.

Results

The best-performing model of the multiple machine learning models assessed was a multilabel logistic regression. The Jaccard coefficient was used for measuring the degree of agreement between a model and a human versus two humans on the same set of codes. The human-to-human agreement consistently outperformed the model-to-human agreement, though both performed best on diagnosis (human-to-human: 0.88, model-to-human: 0.78) and worst on injury codes (human: 0.50, model: 0.28). The model outperformed the human coders on 7.7% of the unique codes assigned by both the model and a human, with strong performance on specific truncated ICD–10–CM diagnosis codes.

Conclusion

This case study demonstrates the potential of machine learning for medical coding in the context of large statistical healthcare surveys. While trained medical coders outperformed the assessed models across the medical coding tasks of assigning correct diagnosis, injury, and RFV codes, machine learning models showed promise in assisting with medical coding projects, particularly if used as an adjunct to human coding.

Keywords: clinical coding • ICD–10–CM • health surveys • NHAMCS–ED • natural language processing

Introduction

Medical coding, the translation of written diagnoses, procedures, and other healthcare information into numeric codes, has traditionally been a difficult and labor-intensive task requiring trained coders to consistently classify medical information into clinically meaningful categories according to the *International Statistical Classification of Diseases and Related Health Problems, 10th Revision, Clinical Modification* (ICD–10–CM), and the National Center for Health Statistics (NCHS) classification for reason for visit (RFV) codes. This study explores the use of machine learning models to replicate medical coding results for the top-level ICD–10–CM and RFV. The objective was to explore the feasibility and effectiveness of machine learning classifiers to perform automated coding of verbatim medical text from patient visits.

Challenges of Medical Coding

Medical coding systems are essential for standardizing complex healthcare processes such as medical payment systems, monitoring use of healthcare services, and tracking public health risks (1). Traditionally, human medical coders assign relevant medical codes by interpreting information in a patient record, including but not limited to case notes, drug charts, and patient administrative data (2). In assigning codes, clinical coders must consider many codes arranged in a hierarchical structure, potentially with multiple codes corresponding to each record in a specific sequence. Electronic dictionary browsers can assist with searches and lookups, but ultimately correct classification relies on the skill of the coder. Coders make judgements despite nonstandard abbreviations, misspellings, and irrelevant information contained in clinical

notes. Though unavoidable, these obstacles make medical coding a difficult and challenging task.

For the Emergency Department portion of the National Hospital Ambulatory Medical Care Survey (NHAMCS–ED), U.S. Census Bureau data collection staff access and abstract patient data from a healthcare provider's records and record these data verbatim into a computerized NHAMCS–ED patient record form. Medical coders do not have access to the full set of information from the patient medical record. Prior research has found that clinical coders with limited patient record information have lower coding accuracy compared with clinical coders with medical support or complete case notes (2). In addition, while medical coders are usually trained on code classification systems that are common for medical billing, some classification systems are used exclusively for healthcare statistical purposes, such as the NCHS RFV classification dictionary. Training medical coders on additional systems and unique coding exceptions can add considerable expense to coding healthcare surveys.

This study seeks to use the high-quality, coded NHAMCS–ED data from NCHS to make the following contributions to the literature on automated medical coding:

1. Assess the use of machine learning for automated coding of patient-level records on both standard medical classification systems (ICD–10–CM) and the RFV classification that is unique to NCHS healthcare surveys.
2. Propose and demonstrate a useful metric for comparing human coders with machine coding (Jaccard similarity) in anticipation of real-world applications of automated coding.

Related Work

Automation of medical coding

Since the advent of electronic clinical information systems, researchers have evaluated opportunities to automate medical coding for a variety of coding classification systems (3). Early machine-automated approaches were deterministic algorithms that assigned codes by matching specific words in a text entry with specific words in a code definition (4). Later, machine-automated approaches expanded the rules-based approach to a larger collection of definitions that included similar terms (5). Some of these approaches also incorporated natural-language processing techniques to handle misspellings, abbreviations, and negation (6). However, these expert-based systems can be laborious to create, extend, and maintain (7,8).

As an alternative, in recent years researchers have applied machine learning to the task of automated coding. Machine learning is a field of artificial intelligence that uses statistical techniques to progressively improve performance on a specific task. Specifically, supervised machine learning methods use observations containing an outcome of interest and various covariates to model how the outcome changes when conditioned on the covariates. This fitted model can then be applied to covariate values of new examples to predict the outcome when it is not present. A primary use of machine learning is for partially or fully automating repetitive, laborious classification tasks currently performed by humans, as machines can generally perform the same repetitive tasks quickly and consistently. This includes the task of classifying text (covariates) as alphanumeric codes (outcomes). In the case of medical coding, a machine learning algorithm can be used to suggest codes for or assign codes to an electronic medical record (EMR) or similar medical text, potentially reducing the time and labor costs of manual medical coding (9).

Several researchers have used machine learning to address the challenge of automating the assignment of alphanumeric codes directly to text in medical documents, including:

- Assignment of ICD–9–CM codes and ICD–9–CM three-digit category to inpatient discharge summaries (10,11)
- Assignment of ICD–9-CM codes to radiology reports (9)
- Assignment of Hospital International Classification of Diseases Adaptation codes, an adaptation of ICD–8 for hospital morbidity, to EMRs (12)
- Assignment of the first digit of ICD–10–CM codes to discharge notes (13)
- Assignment of ICD–10–CM codes to diagnosis descriptions written by physicians in the discharge diagnosis in EMRs (14).

Machine learning models used for these tasks include conventional machine learning classifiers such as K-nearest neighbors (10), Naïve Bayes (11,12), random forests (11), support-vector machines (SVM) (11), logistic regression (11), and example-based classification (12), as well as the growing field of neural networks (9,13,14). No model has emerged as a consistent top performer, though Karimi et al. (9) found that SVM and logistic regression outperformed random forests and that convolutional neural networks could meet or exceed the performance of conventional classifiers for their sample.

It is challenging to develop models that can predict large numbers of unique codes accurately, particularly when some codes occur infrequently (14,15); there are over 70,000 ICD–10 codes (16). To address the large number of codes, researchers have used both truncation of medical codes (9,10,13) and limiting the number of codes evaluated (17). Some researchers have also suggested opportunities to enhance training data using PubMed articles (18).

One key requirement for many machine learning text techniques is a large, high-quality coded dataset (19). These datasets are generally difficult and costly to create, so researchers often resort to using smaller datasets from specific hospitals (2,5,18) or adopting the frequently used and publicly available Medical Information Mart for Intensive Care III (MIMIC–III) dataset (14,20).

Methods for comparing traditional and automated approaches to medical coding

While many automated medical coding models show promise, researchers have raised concerns about validating these models before applying them to real-world tasks (4,19). Because medical codes are often used for financial and policy decisions, accuracy is critical (1). Many researchers evaluate models using precision, recall, and F1 scores on a gold-standard dataset (5,9–11,20). Table 1 shows the definitions of these common evaluation metrics.

While this approach is helpful for comparing performance among models, it does not directly compare model performance with human coder performance. Xu et al. (17) proposed using the Jaccard similarity metric to compare the overlapping text features that physicians and computer models use when assigning ICD–10 codes. Dougherty et al. (21) evaluated humans using automated coding assistance and found an increase in accuracy and a 22% decrease in time per record. Pakhomov et al. (12) implemented an automated medical coding system at the Mayo Clinic where two-thirds of diagnoses were coded with high accuracy, and one-half of these codes did not need to be reviewed manually. The Mayo Clinic was therefore able to reduce the number of staff engaged in manual coding from 34 coders to 7 verifiers.

Methods

Data Source

The datasets used for this evaluation come from the 2016 and 2017 NHAMCS–ED conducted by NCHS. Coding data were drawn from work performed under contract by Research Triangle Institute and its subcontractor HealthCare Resolution Services (HCRS), awarded as "Coding Medical Information in the National Ambulatory Medical Care Survey and the National Hospital Ambulatory Medical Care Survey–Emergency Department." The data used in this evaluation are not available to the public through public-use files but are available for use in the NCHS Research Data Center.

NHAMCS–ED is a nationally representative survey that estimates the use and provision of ambulatory care services in nonfederal, noninstitutional hospital emergency departments (22). The survey is conducted annually and collects information on patient demographics, visits, physician characteristics, and hospital administrative data relevant to healthcare use, healthcare quality, and disparities in healthcare services in the United States.

Patient visit information includes text entries for reason for visit (RFV), cause of injury (CAUSE), and diagnosis (DIAG). Text entries range in length from a phrase to a sentence. Medical coders use the ICD–10–CM for CAUSE and DIAG, except where NCHS instructions supersede standard coding guidelines. Coders use the custom NCHS RFV classification system for RFV. The ICD–10–CM hierarchical classification system uses three to seven characters to denote specific types of morbidity (17). As it encompasses a comprehensive set of codes, it can be used for both classifying diagnosis and cause of injury. The RFV classification system was developed by NCHS to specify reasons for seeking ambulatory medical care and employs a modular, five-character design for each code (23).

In standard survey-coding procedures, medical coders create a sequenced list of codes from a selection of verbatim texts for each variable set (RFV, CAUSE, and DIAG). For each verbatim text field, medical coders may assign zero to several codes, for a total of up to five codes per visit for RFV and DIAG and up to three codes for CAUSE. Every patient record is coded by at least one medical coder. For quality assurance, at least 10% of records are randomly selected for double coding by a second, independent medical coder. If the two coders assign different codes, an adjudicator determines the final code. If disagreement occurs on 5% or more codes, the entire batch is double-coded and adjudicated. These manual coding practices were followed to code the 2016–2017 NHAMCS–ED for this study.

Table 2 provides an overview of the number of visits in the analytic dataset, stratified by the number of codes per visit and the medical code group. The distribution of the number of codes per visit varies across code groups (RFV, CAUSE, and DIAG).

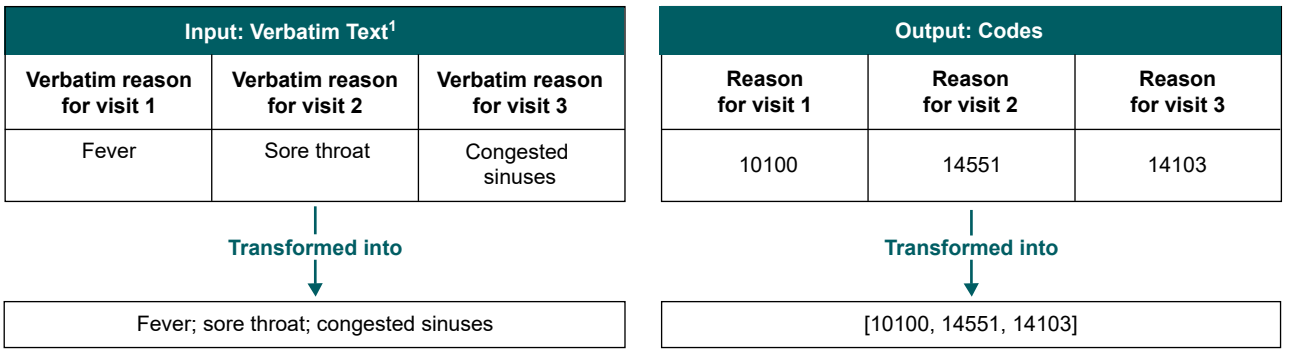
Data Preparation for Machine Learning Analysis

Data preparation is necessary to ensure a high-quality dataset for the purposes of training machine learning models. The dataset included verbatim text entries for DIAG, CAUSE, and RFV, as well as the ICD–10–CM and RFV codes assigned by medical coders for each text entry. In the dataset for machine learning analysis, the record with codes assigned by the initial medical coder was used unless the record had been reviewed by multiple coders and an adjudicated record was available, in which case only the adjudicated record was retained.

The RFV, DIAG, and CAUSE text input could include multiple text fields. To allow for processing by a machine learning model, the multiple verbatim text fields for a patient visit were combined into a single field. Similarly, multiple medical codes were linked together into a single list of codes for each code group. The motivation and validation for this compound approach has been demonstrated by other medical researchers exploring automated coding (12). This process is demonstrated in Figure 1 using a hypothetical RFV example with three verbatim input fields and three associated codes. Combining the verbatim text fields for each visit allowed for the use of a multilabel classification approach.

The ICD–10–CM codes for the CAUSE and DIAG coding tasks were truncated to the three-digit category level. This approach has been previously used by other researchers

Figure 1. Illustration of data element concatenation for three verbatim reasons for visit and three reason-for-visit codes



[1]Variables used in this example are actual survey variables, but the verbatim text and codes did not come from an actual record.
SOURCE: National Center for Health Statistics, National Hospital Ambulatory Medical Care Survey–Emergency Department, 2016–2017.

with ICD–9 codes due to the difficulty of predicting the full unique codes with limited datasets (10,11). Truncating the codes facilitated combining the 2016 and 2017 survey years for CAUSE and DIAG respectively, because categories are generally more stable over time compared with more granular subcategories. In addition, working at the three-digit category level provided a larger number of data records for training and testing each code.

Verbatim text was set to lowercase and transformed using the term frequency–inverse document frequency (TF–IDF) method (24), and patient age and sex were also included in the model. TF–IDF was selected due to the computational limitations of this project. Modern natural-language processing techniques for alternative inputs, such as word embeddings (25) and transformers (26), are promising methods that could enhance future analyses. The data were split using a commonly accepted machine learning heuristic that has been found to be in the range of optimal performance (27), with 80% of the dataset used for developing and training the model (training set) and 20% used for evaluating the out-of-sample model performance (test set).

Multilabel Classification Models

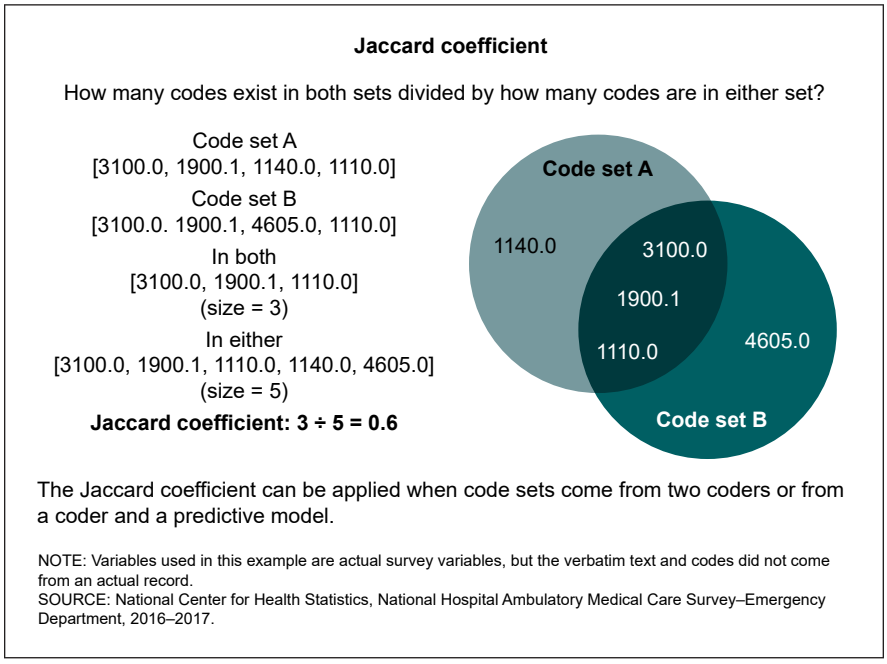
For the standard survey coding, medical coders may assign one or more codes for a given set of verbatim text. To emulate this behavior, the model used a multilabel classification approach. Multilabel text classification models are designed to predict zero to many classes for a given set of text, as opposed to traditional multiclass text classification models, which assign a single class to a set of input text (28). Instead of applying a multiclass model independently for each text field, the modeling paradigm used in this analysis also prevents duplicate code predictions for a specific code group within the same patient visit record.

An evaluation was performed of various multilabel model types in Python using the scikit-learn (29) and scikit-multilearn (30) libraries, including random forests, SVM, and specialized multilabel modeling techniques, such as Multilabel k-Nearest Neighbors. The logistic regression models were both the fastest to train and the most accurate across tasks. For the multilabel context, a composite model was trained consisting of a separate logistic regression model for each code predicting the presence of that code in the input text. To tune the model hyperparameters, a grid search was performed on a separate training set for each case (CAUSE, RFV, DIAG). The resulting model was then trained and evaluated for each case on the initial training and test datasets. The sorted list of predicted codes was truncated by the number of codes that a human coder is permitted to assign (five for RFV and DIAG, three for CAUSE). The best model was selected using the F1 score, the harmonic mean of precision and recall, as the distribution of codes is imbalanced with some codes occurring much more frequently than others and the F1 score is commonly used with imbalanced datasets (31).

Jaccard Coefficient for Comparison to Human Coders

The Jaccard coefficient score was used to compare human medical coders and model predictions to provide more meaningful context for the model results and to understand how the models might perform in a more realistic medical coding setting. The Jaccard coefficient (also referred to as Jaccard similarity) between two sets is the number of items in common divided by the total number of unique items between the two sets (32). The Jaccard coefficient is further detailed in Figure 2. The Jaccard coefficients calculated between 1) the top model predictions and the final medical codes (with the number of model predictions reduced to match the number of final medical codes) were compared against 2) the Jaccard coefficients calculated between independent medical coders for double-coded records.

Figure 2. An explanation and example of the Jaccard coefficient using reason-for-visit codes



NOTE: Variables used in this example are actual survey variables, but the verbatim text and codes did not come from an actual record.
SOURCE: National Center for Health Statistics, National Hospital Ambulatory Medical Care Survey–Emergency Department, 2016–2017.

These double-coded records were part of the quality assurance checks mentioned in the Data Source section. The Jaccard coefficient is not intended to be used as a measure to evaluate coder accuracy or quality, but instead to compare the performance of the model with humans and the relative difficulty of the coding task. Ideally, the Jaccard coefficient between the model predictions and the final medical codes would match or exceed the Jaccard coefficient of the double-coded records, suggesting the model is experiencing similar levels of accuracy and difficulty as a human coder.

To assess how well the model and medical coders performed on individual codes, a modified Jaccard similarity metric was separately calculated for each code. For each code, the following were compared:

- For double-coded records where at least one human coder used the given code, the number of instances when both human coders recorded the code was divided by the total number of times either coder recorded a code; the result ranges from zero to one, where zero occurs when the two human coders never used the code on the same records, and one occurs if the two human coders always used the code on the same records.
- For records with a model prediction and a human assignment, the number of times the model output and the final human assignment both contained the code was calculated and divided by the total number of instances when either the model or human assigned the code; the result ranges from zero to one, where zero occurs when the model and the human never used the code on the same records, and one occurs if the model and the human always used the code on the same records.

For results evaluation, results closer to one are considered particularly strong.

Results

Model Results

In total, three models were trained, one for each code group (RFV, CAUSE, and DIAG). Table 3 shows the precision, recall, and F1-score results for the multilabel classification models in the test sets across the three variable sets. The recall scores are uniformly higher than the precision scores, which indicate that the models make a wide set of predictions that cover the set of true codes relatively well. However, the lower precision values indicate that false positive predictions were common in some modeling tasks. This discrepancy between precision and recall is likely more pronounced in these models because a multilabel classifier can assign multiple codes per verbatim text; with increased predictions, it becomes more likely to detect the correct medical codes at the expense of more false positive predictions.

Comparison to Human Benchmark, by Visit

Table 4 compares the average Jaccard coefficients for the human coders and the model. While the trained human coders consistently outperformed the model in all coding tasks, the human coders and the model showed a similar pattern of performance. Both the human coders and the model did particularly well when assigning truncated ICD–10–CM DIAG codes (humans: 0.88, model: 0.78) and well with RFV codes (humans: 0.83, model: 0.67). Both the model and the humans tended to not perform as well with the truncated ICD–10–CM CAUSE codes (humans: 0.50, model: 0.28).

Comparison to Human Benchmark, by Code

Figures 3–5 illustrate the performance of the model compared with human coders on individual ICD–10–CM and RFV codes. Each dot on the graph reflects a different code. The dot size

corresponds to the number of times a code was used in the dataset by any coder (primary, double coder, or adjudicator). The model-to-human agreement and human-to-human agreement is the Jaccard coefficient, a measure of performance described in Methods. The diagonal reference line represents equivalent performance. For the assigned codes where the model-to-human agreement outperformed human-to-human coder agreement, the dots appear above the line, whereas for the assigned codes where human-to-human coder agreement outperformed the model, the dots are below the line. Note that a code must have been used at least once in the subset of double-coded data and at least once in the test dataset to be considered.

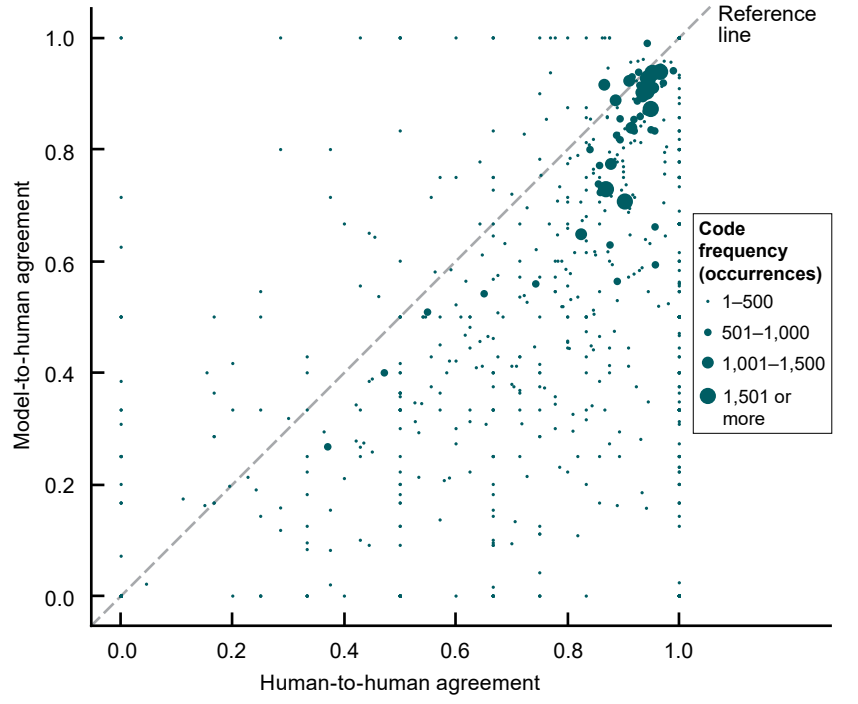
In Figures 3 and 4, many of the larger dots are close to the reference line, illustrating that both the model and the humans performed similarly on frequently used codes. Table 5 provides specific counts for the dots shown in the figures where model agreement exceeded human agreement. Overall, the model-to-human agreement was equal to or greater than the human-to-human agreement for 7.7% of codes. DIAG had the largest proportion of codes with similar or greater performance for model-to-human agreement (10.8%), reflected in the number of dots above the reference line in Figure 3. The model-to-human agreement outperformed the human-to-human agreement on a small percentage of the RFV codes (3.3%), reflected in very few dots above the reference line in Figure 5.

Patterns in performance are further investigated by disaggregating the results with the ICD–10–CM chapter categorizations for DIAG and the RFV module categorizations for RFV in Table 6 and Table 7, respectively. The truncated ICD–10–CM codes for CAUSE only include one ICD–10–CM chapter (External causes of morbidity and mortality), so the summary statistics for CAUSE are the same as in Table 5. In Table 6, the average human-to-human agreement for the DIAG categories is consistently higher than the average model-to-human agreement. The results show a positive trend where human-to-human agreement and model-to-human agreement are both highest on similar ICD–10–CM chapters. There is some positive trend between more occurrences of a chapter and higher levels of agreement, though it is not particularly strong. Table 7 shows similar trends for RFV modules.

Figure 6 and Figure 7 highlight the percentage of codes (truncated in the case of DIAG) in the dataset for each ICD–10–CM chapter or RFV module, respectively, where the model-to-human agreement was higher than the human-to-human agreement. No ICD–10–CM chapter or RFV module resulted in a majority of codes where model-to-human agreement was higher than human-to-human agreement. These findings suggest that the model did not outperform humans when looking at ICD chapters broadly, but it did outperform humans for some specific codes. Four DIAG ICD–10–CM chapters (R00–R99: Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified; J00–J99: Diseases of the respiratory system; Z00–Z99: Factors influencing health status and contact with health services; and P00–P96: Certain conditions originating in the perinatal period) and one RFV module (8: Uncodable entries) had model-to-human agreement greater than human-to-human agreement for approximately 17%–19% of codes in the chapter or module. While the RFV performance may be misleading as it reflects better performance on one code out of six total codes in the Uncodable entries RFV module, the performance on the ICD–10–CM chapters in DIAG are notable as they include a range of unique codes and code frequencies highlighted in Table 6.

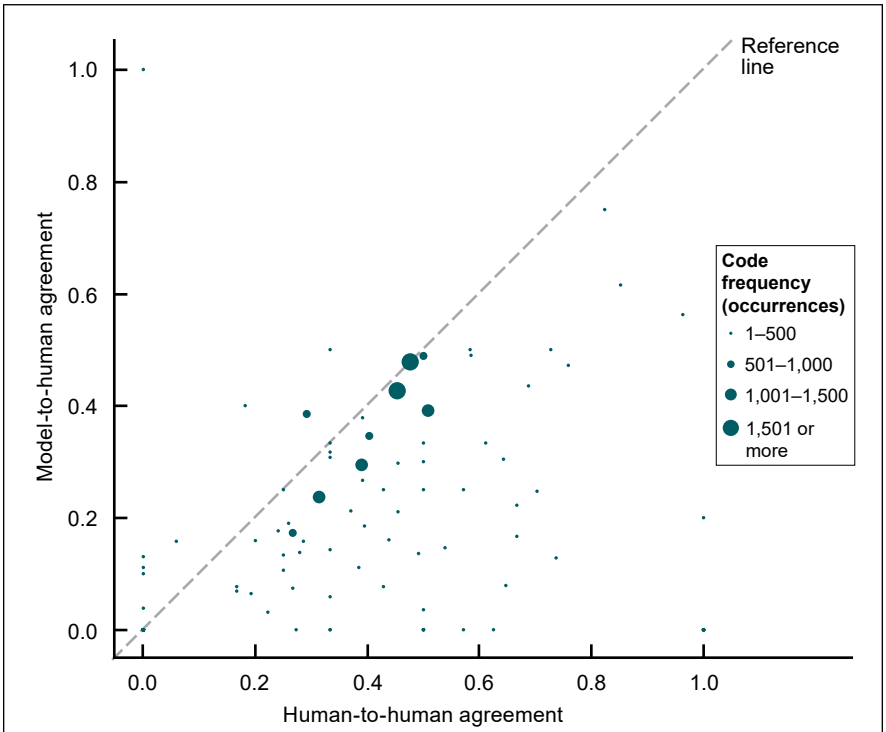
Tables 8–10 list the individual truncated ICD–10–CM codes or the RFV codes where model-to-human agreement was larger than human-to-human agreement. They illustrate the wide variety of codes with stronger model-to-human performance, even within ICD–10–CM chapters and RFV modules. In particular, DIAG codes in Table 8 with large sample sizes that performed well included J02: Acute pharyngitis, J44: Other chronic

Figure 3. Model agreement versus human agreement, by diagnosis



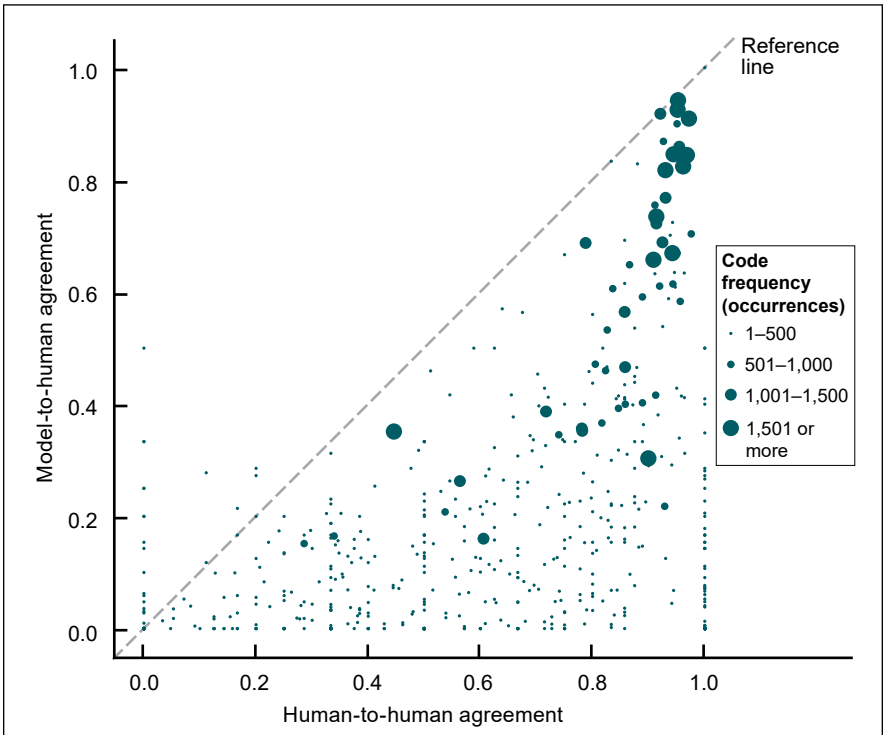
SOURCE: National Center for Health Statistics, National Hospital Ambulatory Medical Care Survey–Emergency Department, 2016–2017.

Figure 4. Model agreement versus human agreement, by cause of injury



SOURCE: National Center for Health Statistics, National Hospital Ambulatory Medical Care Survey–Emergency Department, 2016–2017.

Figure 5. Model agreement versus human agreement, by reason for visit



SOURCE: National Center for Health Statistics, National Hospital Ambulatory Medical Care Survey–Emergency Department, 2016–2017.

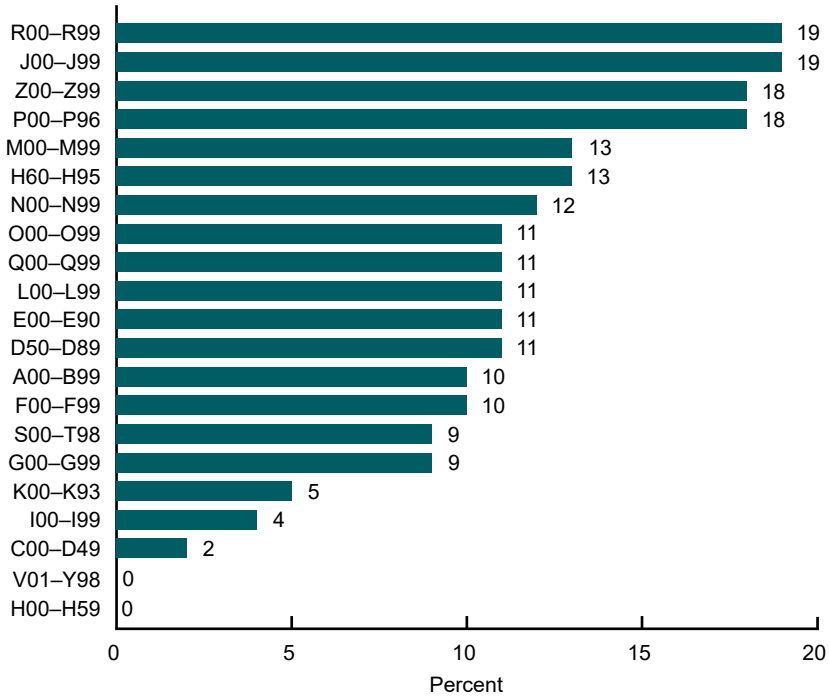
obstructive pulmonary disease, and F17: Nicotine dependence. A number of codes occur infrequently, particularly the RFV codes in Table 10, and Jaccard scores can be low, particularly for CAUSE in Table 9.

Discussion

Though trained human medical coders consistently outperformed the model on the assessed evaluation metrics for the entire code set, the model's results still show promise in approaching human benchmarks, especially for codes used frequently. The results for CAUSE and DIAG add to limited existing literature on automated medical coding for ICD–10 codes, and the high Jaccard coefficients for DIAG may be of particular interest. This project may be the first time that the NCHS RFV classification system has been assessed for automated medical coding, so the results of the RFV analysis contribute novel findings.

A few limited comparisons can be made between this approach and similar approaches in the literature. This study used a multilabel approach and, as expected, the F1 scores were higher than those in the multiclass approach used by Catling et al. (20) with three-digit ICD–9 codes. The recall scores for the top 5 predicted codes for RFV and DIAG exceeded the recall scores for the top 15 three-digit ICD–9 predicted codes using K-nearest-neighbor and Bayesian independence classifiers in Larkey (10), though Larkey used discharge summary data with an average length of 633 words rather than the short text segments used in this project. Logistic regression outperformed random forests and SVM in this project, and Karimi et al. (9) also found that logistic regression outperformed random forests (though not SVM) when predicting 16 unique, full ICD–9 codes in 894 documents. Like other approaches (14,15), the models did not perform as well with the prediction of uncommon codes. Similar to the results in this project, Xu et al. (17) found that the Jaccard coefficient between human coders (physicians) was higher than with a

Figure 6. Percentage of ICD–10–CM codes in data comparing human-to-human and model-to-human Jaccard scores, categorized by ICD-10 chapter

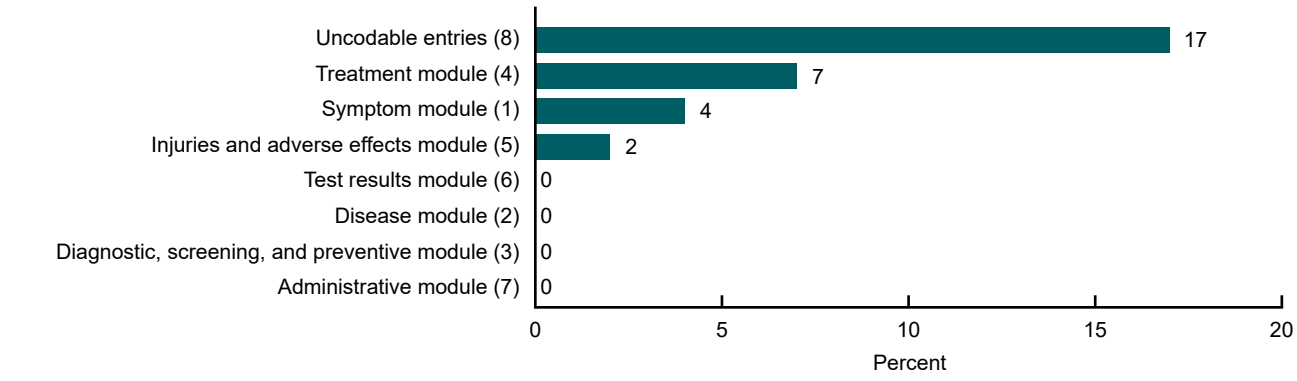


SOURCE: National Center for Health Statistics, National Hospital Ambulatory Medical Care Survey–Emergency Department, 2016–2017.

machine learning model, while also acknowledging that even the human coders did not have perfect agreement. Recognizing that these human and model comparisons can be useful when discussing relative strengths and weaknesses (and similar to the Stanfill et al. [3] review of 113 automated coding studies), the conclusion can be made that, while machine learning holds promise for automated coding tasks, further study is needed to set automated coding performance standards.

There are acknowledged areas for growth in this project and future work could address limitations. Inclusion of more training data could have improved results, especially for infrequent codes and for the CAUSE variables set. Alternatively, limiting the model's results to codes above a certain frequency of use may improve measures of performance. Though existing security and compliance constraints prevented training neural-network models for this data, deep-learning models have shown significant performance improvements on similar modeling tasks (33–35) and could be useful for this application. Incorporating Bayesian hierarchical methods may allow for sequential prediction of codes to better emulate the human medical-coding process (36). Future work could also explore joint inference of code predictions by incorporating information across the coding variable sets (37). In cases where there is new verbatim text that is different than anything previously encountered (for example, a rare disease or condition), the model-predicted probabilities will not fully capture this uncertainty. Advances in high-dimensional inference such as conformal prediction can output more cautious and nuanced predictions, such as a null set ("I don't know") when the text does not resemble the training examples (38).

Figure 7. Percentage of reason-for-visit codes in data comparing human-to-human and model-to-human Jaccard scores, categorized by reason-for-visit module



SOURCE: National Center for Health Statistics, National Hospital Ambulatory Medical Care Survey–Emergency Department, 2016–2017.

Conclusion

While human medical coders still outperform machine learning models in assigning codes to verbatim patient record text, machine learning models show promise in approaching human-level accuracy for coding tasks. Opportunities exist to improve the machine learning models developed in the project and to continue research on appropriate performance metrics for machine learning in automated medical coding.

References

1. American Health Information Management Association. Statement on consistency of healthcare diagnostic and procedural coding. 2007. Available from: https://bok.ahima.org/doc?oid=100524.
2. Tsopra R, Peckham D, Beirne P, Rodger K, Callister M, White H, et al. The impact of three discharge coding methods on the accuracy of diagnostic coding and hospital reimbursement for inpatient medical care. Int J Med Inform 115:35–42. 2018.
3. Stanfill MH, Williams M, Fenton SH, Jenders RA, Hersh WR. A systematic literature review of automated clinical coding and classification systems. J Am Med Inform Assoc 17(6):646–51. 2010.
4. Howell RW, Loy RM. Disease coding by computer. The "fruit machine" method. Br J Prev Soc Med 22(3): 178–81. 1968.
5. Pereira S, Névéol A, Massari P, Joubert M, Darmoni S. Construction of a semi-automated ICD–10 coding help system to optimize medical and economic coding. Stud Health Technol Inform 124:845–50. 2006.
6. Heinze DT, Morsch M, Sheffer R, Jimmink M, Jennings M, Morris W, Morsch A. LifeCode: A deployed application for automated medical coding. AI Magazine 22(2):76. 2001.
7. Farkas R, Szarvas G. Automatic construction of rule-based ICD–9–CM coding systems. BMC Bioinformatics 9(Suppl 3):S10. 2008.
8. Knublauch H. An agile development methodology for knowledge-based systems including a Java framework for knowledge modeling and appropriate tool support. Open Access Repositorium der Universität Ulm und Technischen Hochschule Ulm. 2002. Available from: https://oparu.uni-ulm.de/xmlui/handle/123456789/56.
9. Karimi S, Dai X, Hassanzadeh H, Nguyen A. Automatic diagnosis coding of radiology reports: A comparison of deep learning and conventional classification methods. In: BioNLP. Vancouver, Canada: Association for Computational Linguistics, 328–32. 2017. Available from: https://www.aclweb.org/anthology/W17-2342.
10. Larkey LS, Croft WB. Combining classifiers in text categorization. In: SIGIR '96: Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval. Zurich, Switzerland: Association for Computing Machinery, 289–97. 1996. Available from: https://doi.org/10.1145/243199.243276.
11. Medori J, Fairon C. Machine learning and features selection for semi-automatic ICD–9–CM encoding. In: Louhi '10: Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents. Los Angeles, California: Association for Computational Linguistics, 84–9. 2010.
12. Pakhomov SVS, Buntrock JD, Chute CG. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. J Am Med Inform Assoc 13(5):516–25. 2006.
13. Lin C, Hsu C-J, Lou Y-S, Yeh S-J, Lee C-C, Su S-L, Chen H-C. Artificial intelligence learning semantics via external resources for classifying diagnosis codes in discharge notes. J Med Internet Res 19(11):e380. 2017.
14. Xie P, Xing E. A neural architecture for automated ICD coding. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, 1066–76. Available from: https://www.aclweb.org/anthology/P18-1098.
15. Nguyen AN, Truran D, Kemp M, Koopman B, Conlan D, O'Dwyer J, et al. Computer-assisted diagnostic coding: Effectiveness of an NLP-based approach using SNOMED CT to ICD–10 mappings. AMIA Annu Symp Proc 807–16. 2018.
16. Centers for Disease Control and Prevention. ICD–10–CM official guidelines for coding and reporting. 2019.
17. Xu K, Lam M, Pang J, Gao X, Band C, Mathur P, et al. Multimodal machine learning for automated ICD coding. In: Machine Learning for Healthcare Conference, 197–215. Available from: https://proceedings.mlr.press/v106/xu19a.html.
18. Zhang D, He D, Zhao S, Li L. Enhancing automatic ICD–9–CM code assignment for medical texts with PubMed. In: BioNLP 2017. Vancouver, Canada: Association for Computational Linguistics, 263–71. 2017. Available from: https://www.aclweb.org/anthology/W17-2333.
19. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface 15(141). 2018.
20. Catling F, Spithourakis GP, Riedel S. Towards automated clinical coding. Int J Med Inform 120:50–61. 2018.

21. Dougherty M, Seabold S, White SE. Study reveals hard facts on CAC. J AHIMA 84(7):54–6. 2013.

22. McCaig LF, McLemore T. Plan and operation of the National Hospital Ambulatory Medical Survey. Vital Health Stat 1(34). 1994.

23. Schneider D, Appleton L, McLemore T. A reason for visit classification for ambulatory care. Vital Health Stat 2(78). 1979.

24. Ramos J. Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning, 133–42. 2003.

25. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: Proceedings of Workshop at International Conference on Learning Representations. 2013. Available from: https://arxiv.org/abs/1301.3781.

26. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, et al. Attention is all you need. In: Proceedings of the 31st Conference on Neural Information Processing Systems. 2017. Available from: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

27. Dobbin KK, Simon RM. Optimally splitting cases for training and testing high dimensional classifiers. BMC Med Genomics (4):31. 2011. Available from: https://doi.org/10.1186/1755-8794-4-31.

28. Tsoumakas G, Katakis I. Multi-label classification: An overview. Int J Data Warehous Min. 2007. Available from: https://www.igi-global.com/article/multi-label-classification/1786.

29. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–30. 2011.

30. Szymanski P, Kajdanowicz T. Scikit-multilearn: a scikit-based Python environment for performing multi-label classification. J Mach Learn Res 20(1):209–30. 2019.

31. Sasaki Y. The truth of the F-measure. Teach Tutor Mater. 2007.

32. Levandowsky M, Winter D. Distance between sets. Nature 234:34–5. 1972.

33. Cer D, Yang Y, Kong S, Hua N, Limtiaco N, St John R, et al. Universal sentence encoder. 2018. Available from: https://arxiv.org/abs/1803.11175.

34. Howard J, Ruder S. Universal language model fine-tuning for text classification. 2018. Available from: https://arxiv.org/abs/1801.06146.

35. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. 2018.

36. McCormick TH, Rudin C, Madigan D. Bayesian hierarchical rule modeling for predicting medical conditions. Ann Appl Stat 6(2):652–68. 2012.

37. Zhang Y, Yang Q. A survey on multi-task learning. 2018. Available from: https://arxiv.org/abs/1707.08114.

38. Hechtlinger Y, Póczos B, Wasserman L. Cautious deep learning. 2019. Available from: https://arxiv.org/abs/1805.09460.

Table 1. Classification evaluation metrics

Metric	Definition
Precision	True positives / (true positives + false positives)
Recall	True positives / (true positives + false negatives)
F1 score	Harmonic mean of precision and recall

SOURCE: Japkowicz N, Shah M. Evaluating learning algorithms: A classification perspective. New York, NY: Cambridge University Press. 2011.

Table 2. Number of observations, by the number of codes assigned by medical coders for each code group

Codes per observation	Reason for visit		Cause of injury		Diagnosis	
	Count	Percent	Count	Percent	Count	Percent
At least 1	36,821	100	8,577	100	36,703	100
1	12,826	35	4,954	58	16,287	44
2	9,495	26	2,958	35	9,520	26
3 or more	14,500	39	665[1]	8[1]	10,896	30

[1]Whereas medical coders can code up to five reasons for visit and diagnoses, only three or fewer causes of injury are recorded per visit.

SOURCE: National Center for Health Statistics, National Hospital Ambulatory Medical Care Survey–Emergency Department, 2016–2017.

Table 3. Results from the multilabel classification model for the reason for visit, cause of injury, and diagnosis coding

Code group	Precision	Recall	F1 score
Reason for visit	0.51	0.83	0.60
Cause of injury (3-digit)	0.39	0.70	0.48
Diagnosis (3-digit)	0.74	0.89	0.80

NOTES: Table reports weighted precision, recall, and F1 score. Each metric is calculated for each output code, weighted according to how many visits contained the output code, and averaged.

SOURCE: National Center for Health Statistics, National Hospital Ambulatory Medical Care Survey–Emergency Department, 2016–2017.

Table 4. Comparison in Jaccard coefficients for human-to-human agreement and human-to-model agreement for the reason for visit, cause of injury, and diagnosis coding

Code group	Human-to-human agreement	Human-to-model agreement
Reason for visit	0.83	0.67
Cause of injury (3-digit)	0.50	0.28
Diagnosis (3-digit)	0.88	0.78

SOURCE: National Center for Health Statistics, National Hospital Ambulatory Medical Care Survey–Emergency Department, 2016–2017.

Table 5. Count of codes where model agreement exceeded human agreement for the reason for visit, cause of injury, and diagnosis codes

Dataset	Number of codes model agreement greater than human agreement	Number of considered codes	Number of codes model agreement greater than human agreement as percent of number of considered codes
Reason for visit	22	657	3.3
Cause of injury (3-digit)	10	131	7.6
Diagnosis (3-digit)	97	898	10.8
Total	129	1,686	7.7

SOURCE: National Center for Health Statistics, National Hospital Ambulatory Medical Care Survey–Emergency Department, 2016–2017.

Table 6. Summary of results, by ICD–10–CM chapter for diagnosis-truncated ICD–10–CM codes

ICD–10 chapter name	Jaccard score human-to-human		Jaccard score model-to-human		Truncated codes for chapter in dataset	
	Average	Standard deviation	Average	Standard deviation	Unique	Frequency
Certain conditions originating in the perinatal period P00–P96	0.23	0.410	0.16	0.278	11	76
Certain infectious and parasitic diseases A00–B99	0.75	0.341	0.41	0.362	48	2,431
Congenital malformations, deformations and chromosomal abnormalities Q00–Q99	0.56	0.482	0.09	0.249	18	93
Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism D50–D89	0.64	0.397	0.23	0.335	36	1,132
Diseases of the circulatory system I00–I99	0.81	0.287	0.47	0.341	51	6,148
Diseases of the digestive system K00–K95	0.83	0.224	0.53	0.301	57	5,482
Diseases of the ear and mastoid process H60–H95	0.73	0.326	0.52	0.356	16	1,641
Diseases of the eye and adnexa H00–H59	0.75	0.324	0.45	0.374	19	805
Diseases of the genitourinary system N00–N99	0.77	0.278	0.45	0.320	51	5,180
Diseases of the musculoskeletal system and connective tissue M00–M99	0.72	0.350	0.37	0.366	56	7,233
Diseases of the nervous system G00–G99	0.80	0.272	0.39	0.354	35	2,551
Diseases of the respiratory system J00–J99	0.71	0.290	0.53	0.335	48	8,688
Diseases of the skin and subcutaneous tissue L00–L99	0.66	0.387	0.46	0.375	36	2,812
Endocrine, nutritional and metabolic diseases E00–E89	0.85	0.267	0.53	0.377	28	4,687
External causes of morbidity and mortality V00–Y99	0.00	0.000	0.00	0.000	14	92
Factors influencing health status and contact with health services Z00–Z99	0.55	0.368	0.32	0.294	65	6,597
Injury, poisoning and certain other consequences of external causes S00–T88	0.64	0.305	0.29	0.242	121	15,634
Mental and behavioral disorders F01–F99	0.70	0.338	0.40	0.362	51	6,995
Neoplasms C00–D49	0.76	0.347	0.21	0.235	27	359
Pregnancy, childbirth and the puerperium O00–O9A	0.61	0.380	0.27	0.290	35	1,831
Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified R00–R99	0.75	0.288	0.54	0.338	74	22,290

SOURCE: National Center for Health Statistics, National Hospital Ambulatory Medical Care Survey–Emergency Department, 2016–2017.

Table 7. Summary of results, by reason-for-visit module for reason-for-visit codes

Module	Jaccard score human-to-human		Jaccard score model-to-human		Truncated codes for module in dataset	
	Average	Standard deviation	Average	Standard deviation	Unique	Frequency
1. Symptom module	0.58	0.339	0.18	0.225	395	89,658
2. Disease module	0.67	0.306	0.12	0.210	88	8,887
3. Diagnostic, screening, and preventive module	0.33	0.300	0.08	0.099	23	1,260
4. Treatment module	0.34	0.293	0.10	0.144	45	4,190
5. Injuries and adverse effects module	0.57	0.288	0.14	0.145	89	15,014
6. Test results module	0.54	0.309	0.15	0.149	7	608
7. Administrative module	0.10	0.158	0.04	0.044	4	118
8. Uncodable entries	0.11	0.193	0.01	0.012	6	522

SOURCE: National Center for Health Statistics, National Hospital Ambulatory Medical Care Survey–Emergency Department, 2016–2017.

Table 8. Line listing of truncated ICD–10–CM diagnosis codes comparing human-to-human and model-to-human Jaccard scores, categorized by ICD–10–CM chapter

	ICD–10–CM category and description	Jaccard score human-to-human	Jaccard score model-to-human	Frequency in dataset
	A00–B99: Certain infectious and parasitic diseases			
A08	Viral and other specified intestinal infections	0.70	0.72	123
A31	Infection due to other mycobacteria	0.00	1.00	4
A49	Bacterial infection of unspecified site	0.00	0.31	32
A54	Gonococcal infection	0.00	0.25	7
A69	Other spirochetal infections	0.67	1.00	12
	C00–D49: Neoplasms			
D25	Leiomyoma of uterus	0.38	0.71	47
	D50–D89: Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism			
D62	Acute posthemorrhagic anemia	0.17	0.50	32
D72	Other disorders of white blood cells	0.93	0.96	193
	E00–E90: Endocrine, nutritional and metabolic diseases			
E13	Other specified diabetes mellitus	0.20	0.42	33
E16	Other disorders of pancreatic internal secretion	0.75	0.78	113
E80	Disorders of porphyrin and bilirubin metabolism	0.50	1.00	26
	F00–F99: Mental and behavioral disorders			
F12	Cannabis related disorders	0.91	0.96	141
F13	Sedative, hypnotic, or anxiolytic related disorders	0.50	0.83	40
F17	Nicotine dependence	0.87	0.92	1,023
F41	Other anxiety disorders	0.91	0.92	1,083
F91	Conduct disorders	0.33	0.40	63
	G00–G99: Diseases of the nervous system			
G20	Parkinson's disease	0.88	1.00	42
G62	Other and unspecified polyneuropathies	0.43	0.56	59
G81	Hemiplegia and hemiparesis	0.00	0.17	12
	H60–H95: Diseases of the ear and mastoid process			
H83	Other diseases of inner ear	0.50	1.00	8
H90	Conductive and sensorineural hearing loss	0.00	0.25	8
	I00–I99: Diseases of the circulatory system			
I46	Cardiac arrest	0.75	1.00	57
I70	Atherosclerosis	0.67	0.75	18
	J00–J99: Diseases of the respiratory system			
J00	Acute nasopharyngitis [common cold]	0.67	0.73	71
J02	Acute pharyngitis	0.93	0.94	966
J05	Acute obstructive laryngitis [croup] and epiglottitis	0.86	1.00	167
J09	Influenza due to certain identified influenza viruses	0.64	0.78	71
J10	Influenza due to other identified influenza virus	0.44	0.65	82
J38	Diseases of vocal cords and larynx, not elsewhere classified	0.29	1.00	11
J43	Emphysema	0.60	1.00	32
J44	Other chronic obstructive pulmonary disease	0.94	0.99	759
J80	Acute respiratory distress syndrome	0.00	0.71	23
	K00–K93: Diseases of the digestive system			
K03	Other diseases of hard tissues of teeth	0.60	0.75	19
K31	Other diseases of stomach and duodenum	0.67	0.83	51
K37	Unspecified appendicitis	0.43	1.00	24
	L00–L99: Diseases of the skin and subcutaneous tissue			
L22	Diaper dermatitis	0.83	0.88	48
L24	Irritant contact dermatitis	0.00	0.50	7
L60	Nail disorders	0.80	1.00	32
L93	Lupus erythematosus	0.00	0.25	5

Table 8. Line listing of truncated ICD–10–CM diagnosis codes comparing human-to-human and model-to-human Jaccard scores, categorized by ICD–10–CM chapter—Con.

	ICD–10–CM category and description	Jaccard score human-to-human	Jaccard score model-to-human	Frequency in dataset
	M00–M99: Diseases of the musculoskeletal system and connective tissue			
M06	Other rheumatoid arthritis	0.77	1.00	60
M10	Gout	0.75	0.90	128
M21	Other acquired deformities of limbs	0.00	1.00	9
M32	Systemic lupus erythematosus (SLE)	0.83	0.86	41
M53	Other and unspecified dorsopathies, not elsewhere classified	0.57	0.75	28
M65	Synovitis and tenosynovitis	0.25	0.50	29
M94	Other disorders of cartilage	0.87	1.00	86
	N00–N99: Diseases of the genitourinary system			
N12	Tubulo-interstitial nephritis, not specified as acute or chronic	0.56	0.71	95
N30	Cystitis	0.87	0.95	302
N31	Neuromuscular dysfunction of bladder, not elsewhere classified	0.00	0.50	9
N52	Male erectile dysfunction	0.50	1.00	5
N61	Inflammatory disorders of breast	0.33	0.43	21
N95	Menopausal and other perimenopausal disorders	0.67	1.00	11
	O00–O99: Pregnancy, childbirth and the puerperium			
O42	Premature rupture of membranes	0.00	0.63	18
O47	False labor	0.45	0.64	67
O62	Abnormalities of forces of labor	0.17	0.36	25
O80	Encounter for full-term uncomplicated delivery	0.25	0.33	26
	P00–P96: Certain conditions originating in the perinatal period			
P28	Other respiratory conditions originating in the perinatal period	0.00	0.50	4
P59	Neonatal jaundice from other and unspecified causes	0.50	0.60	20
	Q00–Q99: Congenital malformations, deformations and chromosomal abnormalities			
Q05	Spina bifida	0.50	1.00	7
Q24	Other congenital malformations of heart	0.00	0.17	10
	R00–R99: Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified			
R03	Abnormal blood-pressure reading, without diagnosis	0.76	0.85	255
R04	Hemorrhage from respiratory passages	0.94	0.96	198
R05	Cough	0.92	0.93	922
R06	Abnormalities of breathing	0.89	0.89	1,065
R14	Flatulence and related conditions	0.83	1.00	46
R44	Other symptoms and signs involving general sensations and perceptions	0.38	0.80	49
R46	Symptoms and signs involving appearance and behavior	0.20	0.33	18
R47	Speech disturbances, not elsewhere classified	0.67	0.75	37
R52	Pain, unspecified	0.15	0.16	181
R73	Elevated blood glucose level	0.84	0.85	277
R74	Abnormal serum enzyme levels	0.64	0.67	92
R76	Other abnormal immunological findings in serum	0.00	0.33	10
R78	Findings of drugs and other substances, not normally found in blood	0.17	0.29	20
R99	Ill-defined and unknown cause of mortality	0.67	1.00	12
	S00–T98: Injury, poisoning and certain other consequences of external causes			
S13	Dislocation and sprain of joints and ligaments at neck level	0.56	0.58	138
S67	Crushing injury of wrist, hand and fingers	0.80	0.88	40
S68	Traumatic amputation of wrist, hand and fingers	0.40	0.67	25
S72	Fracture of femur	0.77	0.94	103
T21	Burn and corrosion of trunk	0.00	0.33	24
T31	Burns classified according to extent of body surface involved	0.00	0.38	22
T51	Toxic effect of alcohol	0.11	0.17	36
T65	Toxic effect of other and unspecified substances	0.17	0.29	23
T68	Hypothermia	0.00	0.50	7
T76	Adult and child abuse, neglect and other maltreatment, suspected	0.30	0.32	62
T80	Complications following infusion, transfusion and therapeutic injection	0.00	0.17	16

Table 8. Line listing of truncated ICD–10–CM diagnosis codes comparing human-to-human and model-to-human Jaccard scores, categorized by ICD–10–CM chapter—Con.

ICD–10–CM category and description		Jaccard score human-to-human	Jaccard score model-to-human	Frequency in dataset
	Z00–Z99: Factors influencing health status and contact with health services			
Z00	Encounter for general examination without complaint, suspected or reported diagnosis	0.46	0.54	187
Z09	Encounter for follow-up examination after completed treatment for conditions other than malignant neoplasm	0.00	0.07	46
Z37	Outcome of delivery	0.25	0.55	37
Z46	Encounter for fitting and adjustment of other devices	0.15	0.40	53
Z47	Orthopedic aftercare	0.29	0.80	19
Z53	Persons encountering health services for specific procedures and treatment, not carried out	0.19	0.20	129
Z59	Problems related to housing and economic circumstances	0.88	1.00	76
Z63	Other problems related to primary support group, including family circumstances	0.00	0.20	21
Z65	Problems related to other psychosocial circumstances	0.00	0.25	28
Z72	Problems related to lifestyle	0.75	0.75	398
Z76	Persons encountering health services in other circumstances	0.72	0.83	195
Z94	Transplanted organ and tissue status	0.78	1.00	44

SOURCE: National Center for Health Statistics, National Hospital Ambulatory Medical Care Survey–Emergency Department, 2016–2017.

Table 9. Line listing of truncated ICD–10–CM cause-of-injury codes comparing human-to-human and model-to-human Jaccard scores, categorized by ICD–10–CM chapter

ICD–10–CM category and description		Jaccard score human-to-human	Jaccard score model-to-human	Frequency in dataset
	V01–Y98: External causes of morbidity and mortality			
V49	Car occupant injured in other and unspecified transport accidents	0.18	0.40	465
V74	Bus occupant injured in collision with heavy transport vehicle or bus	0.00	1.00	3
V89	Motor- or nonmotor-vehicle accident, type of vehicle unspecified	0.29	0.39	576
W05	Fall from non-moving wheelchair, nonmotorized scooter and motorized mobility scooter	0.33	0.50	46
W51	Accidental striking against or bumped into by another person	0.06	0.16	96
X10	Contact with hot drinks, food, fats and cooking oils	0.00	0.10	18
X15	Contact with hot household appliances	0.00	0.11	16
Y07	Perpetrator of assault, maltreatment and neglect	0.00	0.13	47
Y84	Other medical procedures as the cause of abnormal reaction of the patient, or of later complication, without mention of misadventure at the time of the procedure	0.00	0.04	36
Y92	Place of occurrence of the external cause	0.48	0.48	2,759

SOURCE: National Center for Health Statistics, National Hospital Ambulatory Medical Care Survey–Emergency Department, 2016–2017.

Table 10. Line listing of reason-for-visit codes comparing human-to-human and model-to-human Jaccard scores, categorized by module

Reason-for-visit code	RFV code description	Jaccard score human-to-human	Jaccard score model-to-human	Frequency in dataset
	1: Symptom module			
1080.3	Feeding problem, eating difficulty	0.00	0.06	20
1135.0	Disturbances of sleep	0.00	0.03	33
1150.1	Substance abuse, not otherwise specified	0.00	0.15	24
1455.6	Lump or mass	0.11	0.28	45
1470.0	Abnormalities of sputum or phlegm	0.00	0.05	25
1515.4	Abnormal color, ridges, coated	0.00	0.50	8
1575.0	Difficulty eating	0.20	0.27	36
1610.0	Symptoms of liver, gallbladder, and biliary tract	0.00	0.14	13
1665.0	Symptoms of bladder	0.20	0.29	28
1670.0	Symptoms of the kidneys	0.00	0.20	10
1740.3	Abnormal material, including clots	0.00	0.10	22
1895.0	Navel problems	0.00	0.04	22
1925.0	Knee symptoms	0.11	0.12	44
1930.4	Ankle symptoms—weakness	0.00	0.33	6
1935.3	Foot and toe symptoms—limitation of movement, stiffness, tightness	0.00	0.33	7
1935.4	Foot and toe symptoms—weakness	0.00	0.20	11
	4: Treatment module			
4415.0	Radiation therapy	0.00	0.25	14
4518.0	Detoxification	0.17	0.21	49
4520.0	Minor surgery	0.00	0.06	22
	5: Injuries and adverse effects module			
5050.0	Fracture, other and unspecified	0.00	0.25	8
5830.1	Sexual abuse	0.00	0.03	14
	8: Uncodable entries			
8999.0	Illegible entry	0.00	0.01	42

SOURCE: National Center for Health Statistics, National Hospital Ambulatory Medical Care Survey–Emergency Department, 2016–2017.

Vital and Health Statistics Series Descriptions

Active Series

Series 1. **Programs and Collection Procedures**
Reports describe the programs and data systems of the National Center for Health Statistics, and the data collection and survey methods used. Series 1 reports also include definitions, survey design, estimation, and other material necessary for understanding and analyzing the data.

Series 2. **Data Evaluation and Methods Research**
Reports present new statistical methodology including experimental tests of new survey methods, studies of vital and health statistics collection methods, new analytical techniques, objective evaluations of reliability of collected data, and contributions to statistical theory. Reports also include comparison of U.S. methodology with those of other countries.

Series 3. **Analytical and Epidemiological Studies**
Reports present data analyses, epidemiological studies, and descriptive statistics based on national surveys and data systems. As of 2015, Series 3 includes reports that would have previously been published in Series 5, 10–15, and 20–23.

Discontinued Series

Series 4. **Documents and Committee Reports**
Reports contain findings of major committees concerned with vital and health statistics and documents. The last Series 4 report was published in 2002; these are now included in Series 2 or another appropriate series.

Series 5. **International Vital and Health Statistics Reports**
Reports present analytical and descriptive comparisons of U.S. vital and health statistics with those of other countries. The last Series 5 report was published in 2003; these are now included in Series 3 or another appropriate series.

Series 6. **Cognition and Survey Measurement**
Reports use methods of cognitive science to design, evaluate, and test survey instruments. The last Series 6 report was published in 1999; these are now included in Series 2.

Series 10. **Data From the National Health Interview Survey**
Reports present statistics on illness; accidental injuries; disability; use of hospital, medical, dental, and other services; and other health-related topics. As of 2015, these are included in Series 3.

Series 11. **Data From the National Health Examination Survey, the National Health and Nutrition Examination Surveys, and the Hispanic Health and Nutrition Examination Survey**
Reports present 1) estimates of the medically defined prevalence of specific diseases in the United States and the distribution of the population with respect to physical, physiological, and psychological characteristics and 2) analysis of relationships among the various measurements. As of 2015, these are included in Series 3.

Series 12. **Data From the Institutionalized Population Surveys**
The last Series 12 report was published in 1974; these reports were included in Series 13, and as of 2015 are in Series 3.

Series 13. **Data From the National Health Care Survey**
Reports present statistics on health resources and use of health care resources based on data collected from health care providers and provider records. As of 2015, these reports are included in Series 3.

Series 14. **Data on Health Resources: Manpower and Facilities**
The last Series 14 report was published in 1989; these reports were included in Series 13, and are now included in Series 3.

Series 15. **Data From Special Surveys**
Reports contain statistics on health and health-related topics from surveys that are not a part of the continuing data systems of the National Center for Health Statistics. The last Series 15 report was published in 2002; these reports are now included in Series 3.

Series 16. **Compilations of Advance Data From Vital and Health Statistics**
The last Series 16 report was published in 1996. All reports are available online; compilations are no longer needed.

Series 20. **Data on Mortality**
Reports include analyses by cause of death and demographic variables, and geographic and trend analyses. The last Series 20 report was published in 2007; these reports are now included in Series 3.

Series 21. **Data on Natality, Marriage, and Divorce**
Reports include analyses by health and demographic variables, and geographic and trend analyses. The last Series 21 report was published in 2006; these reports are now included in Series 3.

Series 22. **Data From the National Mortality and Natality Surveys**
The last Series 22 report was published in 1973. Reports from sample surveys of vital records were included in Series 20 or 21, and are now included in Series 3.

Series 23. **Data From the National Survey of Family Growth**
Reports contain statistics on factors that affect birth rates, factors affecting the formation and dissolution of families, and behavior related to the risk of HIV and other sexually transmitted diseases. The last Series 23 report was published in 2011; these reports are now included in Series 3.

Series 24. **Compilations of Data on Natality, Mortality, Marriage, and Divorce**
The last Series 24 report was published in 1996. All reports are available online; compilations are no longer needed.

For answers to questions about this report or for a list of reports published in these series, contact:

Information Dissemination Staff
National Center for Health Statistics
Centers for Disease Control and Prevention
3311 Toledo Road, Room 4551, MS P08
Hyattsville, MD 20782

Tel: 1–800–CDC–INFO (1–800–232–4636)
TTY: 1–888–232–6348
Internet: https://www.cdc.gov/nchs
Online request form: https://www.cdc.gov/info
For e-mail updates on NCHS publication releases, subscribe online at: https://www.cdc.gov/nchs/email-updates.htm.

U.S. DEPARTMENT OF HEALTH & HUMAN SERVICES

Centers for Disease Control and Prevention
National Center for Health Statistics
3311 Toledo Road, Room 4551, MS P08
Hyattsville, MD 20782–2064

OFFICIAL BUSINESS
PENALTY FOR PRIVATE USE, $300

FIRST CLASS MAIL
POSTAGE & FEES PAID
CDC/NCHS
PERMIT NO. G-284

For more NCHS Series Reports, visit:
https://www.cdc.gov/nchs/products/series.htm

Series 2, No. 189
CS326148