BUILDING A LARGE PUBLIC DATABASE OF DE-IDENTIFIED ELECTROCARDIOGRAM DATA FROM A SINGLE EMERGENCY DEPARTMENT

By

David Yale Gelman, MD

A CAPSTONE PROJECT

Presented to the Department of Medical Informatics and Clinical Epidemiology
School of Medicine
in partial fulfillment of
the requirements for the degree of

Master of Biomedical Informatics

September 2017

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Master's Capstone of

David Yale Gelman, MD

has been approved

[redacted]

William Hersh, MD

Acknowledgements

I wish to thank Dr. Martin Pusic for sponsoring this master's capstone project. He has been very welcoming of me into the NYU Langone family as well as providing insight and guidance. Dr. William Hersh at OHSU has been an inspiration to the informatics community and has personally inspired me since I first met him at a 10x10 course at OHSU. I should also credit the ARRA HITECH Act for covering the graduate certificate portion of my training. Additionally, I'd like to thank Dr. Curt Dill, the chief of the emergency department in which I work at the Manhattan Veterans Hospital for allowing my participation in this project and for supporting my graduate education. And most importantly, I am endlessly indebted to my devoted husband Eric Blume for his boundless support as well as for putting up with all of my endless prattling about informatics.

I dedicate this work to my parents, Arnold and Janice Gelman, who have always encouraged me to pursue my interests, cultivated in me a sense of wonder and a deep interest in learning. I love them immeasurably.

Table of Contents

Abstract

This summary describes the creation of a publically-available database of over 80,000 electrocardiogram (ECG) tracings obtained from patients treated in a single emergency department in a three year period. It describes the process of obtaining the files from the clinical ECG database, processing the text of the electrocardiogram reads, and classifying the findings within the ECGs against a preexisting terminology standard. Future work includes using the qualitative method of thematic analysis to identify categories of concepts not expressed fully in the terminology standard. The issues and methodology of de-identification to prepare the database for public release are discussed.

Keywords: classification, de-identification, emergency medicine, electrocardiogram

Introduction

Many of the electrocardiograms (ECGs) performed on patients are recorded and stored in an electronic format. In spite of this, there are very few if any large databases of de-identified electrocardiographic data. Results of an incomplete review of resources discovered via online searches of web sites and publications are located in Appendix A. The largest database found was a Brazilian telecardiology site with 1.9 million clinically-obtained ECGs, but the records are not publically available. (Chazard et al., 2015) Other resources tend to have lower numbers of participants or records; these often serve as reference materials for patients with specific clinical conditions. Many of these data sets include only ECGs or cardiac waveform tracings captured in controlled settings which are often used as standards against which to test medical equipment. This work seeks to create a resource that was not found to exist – a large, de-identified database of ECGs with findings classified against a terminology standard. Our data set happens to be collected from a single emergency department in a large urban tertiary-care hospital.

This capstone is a subproject within a currently-funded Department of Defense (DOD) grant. Dr. Martin Pusic of the New York University (NYU) School of Medicine Institute for Innovations in Medical Education obtained a grant from the DOD Medical Simulation and Information Sciences Research Program to develop an adaptive tutor for improving the visual recognition and diagnosis of findings within electrocardiograms. Dr. Pusic is working in collaboration with colleagues from the Mayo Clinic, the University of British Columbia and the University of Illinois (Chicago). The grant ID is W81XWH-15-DMRDP-MSIS-ATUMN and its planned duration extends from 10/2016 until 9/2018.

The DOD sought researchers interested in developing a system in which a medical trainee could improve a skill in a domain of visual medical diagnosis with a computer system supervising the training. Dr. Pusic's accepted proposal was to collect and classify 20,000 consecutive ECGs performed in the NYU Langone Medical Center (NYULMC) emergency department. These ECGs would be classified into a formal structure of categories which represent the findings present on the cardiograms. A graphic-user interface would be designed that would allow a student to be shown ECG waveforms and the student would have to input their assessment of the findings present. The system would score a student's responses and present new material in a supervised manner, continually assessing their level of mastery at identifying the wide variety of potential findings within ECGs. Students would be challenged with ECGs of increasing complexity, ramping very quickly to their current skill level. The system would also retest learners on material considered already mastered to ensure persistence of knowledge. The computer-aided teaching tool would be compared to the usual and customary practice of instructor-led ECG training to assess its benefit.

Dr. Pusic's prior work has focused on understanding the path students take on achieving mastery of subjects in medical education and using knowledge of the various patterns of learning curves to design better teaching systems. He has evaluated pediatric emergency medicine resident's learning curves during a deliberate learning process classifying ankle x-rays into normal or abnormal. (M. Pusic, Pecaric, & Boutis, 2011) The study involved the design of a computer program to display the x-rays and the study patient's chief complaint and clinical findings. They collected the learner's response, which included marking the location of abnormality on the image for those images interpreted as abnormal, captured the data on learner's responses, and presented immediate feedback to the learner regarding the correctness of their response. Thestudy was fairly unique at the time in that it assessed the pattern of learning over time and rather than intermittent assessment which merely serves as a snapshot of a learner's current knowledge. In follow-up, Pusic and colleagues also discovered that altering the frequency of abnormal cases in a training set can impact the sensitivity and specificity of the learner in interpretation of the findings in a dose-response pattern. A greater percentage of negative cases in the test material led to increased accuracy by the learners with negative x-rays and thus a higher specificity for this task. Those exposed to greater percentage of positives in the training cases led to learners with higher sensitivity for identifying fractures. (M. V. Pusic et al., 2012)

This grant seeks to expand on his prior work. Rather than limiting learners to deciding on the presence or absence of a single finding (normal vs. abnormal x-rays), learners will be operating on a domain that is multifactorial; each ECG can contain or not contain any of nearly 120 findings. Assessing a student's learning curve across the full domain of ECG interpretation is a much more complicated endeavor. The ankle x-ray studies required a collection of de-identified ankle x-rays correctly identified as normal or abnormal. This master's capstone project seeks to develop a collection of ECGs that are classified appropriately and arranged into a database that can supply the computerized training system.

In addition, the database will be de-identified for eventual public release. The increasing interest in large public datasets drove the decision to create a de-identified version of our data.

Methods

Obtaining study data

A flow chart describing the data processing workflow is available in Appendix B. A dedicated, shared digital storage space drive behind the NYU firewall was established for collaborative work on this project. Approval was sought and obtained for collection of data for this project through the IRB of NYU Medical Center. Two primary sources of data were accessed – the General Electric MUSE Cardiology System and the Epic Clarity clinical data warehouse. (Integration of the clinical data from the Epic EHR was beyond the scope of the capstone project). The initial dataset was comprised of 98,420 ECGs obtained from the MUSE system. These records were limited to patients treated in the Perelman Department of Emergency Medicine of NYU Langone Medical Center over a roughly 5 year period. The exact time span or actual start and end dates were not revealed here as this would potentially make it easier to re-identify patients from the data. The MUSE database search also limited the obtained ECGs to those formally interpreted by a cardiologist. As per the DOD study protocol, ECGs of patients younger than 18 or older than 80 years were excluded from this initial dataset. Fifteen cases were excluded for data quality issues in the structure of the ECG data files. A total of 81,287 resting electrocardiograms remained.

Investigating the structure of the MUSE ECG data file

Each electrocardiogram record from MUSE consists of an XML-formatted file. This file contains many sections of data, but the four main sections are patient demographics, study demographics, text interpretation both from the MUSE analysis software and from the reading cardiologist, and waveform data. The data in the patient demographics and study demographics sections was fairly simple to gather into a comma-separated value (CSV) file as all of the entries in these two sections consisted of a single data item per heading per electrocardiogram. In the initial version of this data, all elements including personally identifiable information (PII) were preserved. A Python programming language script was written which extracted the various elements from each XML into a single CSV document; these would serve as the raw materials for building a Structured Query Language (SQL) database.

The MUSE XML data contains the text of the findings found within the ECG as interpreted by the MUSE analysis software. The so-called “machine read” is the text that the MUSE system generates when the ECG is uploaded to its database. It is presented, in the MUSE XML data as a series of text phrases with the XML header of “OriginalDiagnosis”. Each phrase is also labeled as to whether that specific phrase constituted the end line of text for the original ECG printout. Also within the XML data is the text of the cardiologist’s edited and approved reading, labeled as individual phrases under the XML heading of “Diagnosis”. A Python programming language script was written which stitched together the individual phrases within each of these two categories of reads into individual lines of text. These lines of text served as the raw material for the classification system, described in subsection “Classifying the parsed files into categories” below.

Interpreting the waveform data was not specifically within the scope of the capstone project. However, noted here for reference, a rather useful paper was identified which helped the research team reverse engineer the waveform data stored in the MUSE XML files into actual time series data of the millivolt recordings of each cardiac lead. (Popa, 2011)

Applying text processing to the MUSE ECG XML Files

The goal of the first stage of data processing was to convert the initial XML files into Python data objects, process the text in the files, and output to JSON format. Data in the PatientDemographics, TestDemographics and Waveform XML-headed sections of the ECG were not touched during this process – they remained intact in structure and content in the resulting JSON files. Only the OriginalDiagnosis and Diagnosis XML-headed sections underwent processing; these sections contain the text of the machine interpretation and the cardiologist's final interpretation respectively. An example of initial XML file text and the results of the various steps in this process are in Appendix C.

To convert from XML to JSON, each subsection with the DiagnosisStatement needs to contain the text of the statement (like "Anteroseptal infarct" or "Abnormal ECG"), an endline flag set to True or False, and a user insert flag which is set to True or False. The endline flag denotes that this statement ends a specific line of text in the reading, and the user insert flag denotes that the statement was added manually during review by the cardiologist and was not part of the original machine interpretation.

The processor then concatenates all of the individual diagnosis statements into single lines of text with the cutoff delineated by the endlineFlag. Subsequently, it uses regular expressions to identify text lines that contain text that denote that a comparison between current and prior ECGs is being made. When it finds these statements, it changes a comparisonFlag to True. This was done an assist to future work that might be aided by the capture of whether a given ECGs contains text referring to a prior ECGs.

The text within the ECG interpretations was stripped of all dates. However, it was important to maintain the meaning or context that these dates provided. Almost all of the dates in the ECG interpretations were included as references to prior studies. To preserve the meaning, the dates were substituted with text describing the time interval in question. For instance, if an ECG was obtained on 1/1/2010 and contained the text line "compared to an ECG of 6/7/2010", the Python script would convert this text to read "compared to an ECG 5 months ago". The text interpretations within the ECG data were fairly consistent in its use of dates. Dates were found only in specific text phrases. This made conversion much easier.

The last step of the initial process was to write the in-memory Python dictionary objects to JSON format. New headers were created to incorporate data generated from the processing steps, such as the concatenated text of each line of the interpretations. The text lines that were the result of concatenation were saved to a separate data element to preserve the original data. The result of this initial process was a folder containing JSON files in a one-to-one relationship with the original files.

Classifying the parsed files into categories

Choosing the standardized terminology to use for classification of ECGs. There is no International Standards Organization (ISO) terminology standard for the classification of ECG findings. There is an ISO standard for how medical waveform data should be encoded within ECGs, ISO/TS 22077-2:201. Additionally, ISO 11073-91064:2009 sets a standard for the communication protocol or transmitting ECG data between devices. There is also an HL7 aECG standard, but it doesn't restrict that findings be represented using a standard terminology; they are recorded as text only. (Brown, 2005)

An initial terminology standard for findings within ECGs has been established via a consensus statement from the various expert panels in the field of cardiology. (Mason et al., 2007) An incomplete list of terms in this standard is available in Appendix D for example. The terminology consists of 117 “primary diagnostic statements” arranged into categories. In addition, there are additional modifier terms to alter the meaning of the primary statements. Also, there are “secondary statements” which are often disease states that findings within a given ECG might suggest or for a clinician to consider, but because the findings are not pathognomonic for the condition, they cannot be primary statements in and of themselves.

A few terms were added to the list compiled by Mason et al. above. These terms were either found to be very common in the text of readings and not represented in the terminology. These included “normal axis”, “repolarization abnormality”, “late transition” and “compared with prior ECG,” and “abnormal qrs-t angle, consider primary t-wave abnormality”. Also added was “Wolff-Parkinson-White” – although not common, it was a specific condition worth capturing under its commonly-used name.

Choosing a classification method. There are many different techniques for classifying text. They encompass brute force methods, Naïve Bayes algorithms, k-means clustering, neural networks, and more. This work required that text be classified into more than one category if it exists, so binary classifiers were not an appropriate choice for this project. Two specific factors were heavily weighted in the decision of what technique to use. The first was that the data is highly structured – the ECG analysis software has a limited vocabulary and the NYU cardiologists rely heavily on use of canned or patterned text (both native to the MUSE system and locally-configured terms). The second and most important was that the output of the classifier was going to be treated as the correct interpretation for the findings present within the ECG when students were engaged with the ECG teaching system. The classified text is meant to be a one-to-one exact representation to the machine and cardiologist readings. Because of these two factors, it was decided that hand classification by an expert would be the ideal method.

Once this was decided, it followed that the classifier should process the readings on a line-by-line basis, and not by individual words or phrases. The findings in an ECG reading are more easily understood and identified in a line of text than in a word or phrase. Of critical importance is the problem of negation; many ECGs have text which compares the current ECG to the prior of the same patient and comments on features which were new or which were no longer present in the prior. For example, “Compared to an ECG of 6 weeks ago, atrial fibrillation has replaced sinus rhythm.” Classifying by phrase might fail to recognize that sinus rhythm is no longer present in the current ECG leading to a false positive match on this term.

Preparing the classifier dictionary. A brief Python script was designed to collect each unique like of text from every machine and cardiology reading in the processed JSON files. A total of approximately 3,500 unique lines of text were obtained. Given that the data set contains nearly 100,000 ECGs of a median of 4 lines each, the relative paucity of number of unique lines of text was surprising and spoke to the structured nature of the data. These 3,500 lines were classified by hand, comparing the text in the lines to the terms in the terminology standard. The lines of text were placed in one column of a database. If the line described a finding as being present, the number corresponding to the term was added to a second column of current findings. (Multiple findings were comma-separated in a given column.) In a subtle distinction, if the line described a comparison to an old ECG with a new finding present that wasn’t there before, the numerical ID value of that term was placed in a separate column of new findings. A third column for findings no longer present was maintained, for findings present on prior ECGs that were not evidenced in the current. An ECG reading comprised of the lines “Normal sinus rhythm, compared to a prior ECG of 2 months ago, right bundle branch block has replaced incomplete right bundle branch block” would be classified into current finding of 20 (for normal sinus rhythm), new finding of 106 (for newly-appeared right bundle branch block) and prior finding of 105 (for the incomplete right bundle, identified as prior because it was no longer present).

As with any attempt at matching text to a terminology, there were some clinical ideas or phrases which did not have a clear equivalent within the terminology. For instance, the terminology doesn’t have specific terms to identify cardiac regions (anterior, inferior, etc) or specific ECG leads (I, aVL, V1, etc.). In other cases, a clinical term could match to more than one term or at least, by text alone, it was ambiguous to what specific term it would match. An example of this is “second degree heart block”. This could match to “Second-degree AV block, Mobitz type I (Wenckebach)” as item 82 or it could match to “Second-degree AV block, Mobitz type II” as item 83. This distinction would have to be resolved by actual reevaluation of the EKG tracing.

Classifying the ECG findings into standardized terminology. A separate Python script was created which loads each ECG JSON file, compares the individual lines of text of the machine and cardiology interpretations to the dictionary of 3,500 classified lines, and saves the AHA classes which correlate with each line back to the JSON file.

Results

Assessing the accuracy of the classification process

The initial idea was to use a panel of experts (board-certified cardiologists) to assess the degree of agreement between the original text of ECG readings with the output of the classifier. They would rate the degree of agreement on a visual analog scale with 1 representing no agreement whatsoever between the two sources and 7 representing full agreement (preservation of meaning) from the original to the classified interpretation of the ECG. A Cohen's weighted kappa score was used to assess the interrater validity. (Cohen, 1968) However, this approach is deficient in one significant aspect – there is no visual analog scale that has been specifically validated for this task. Creating a scale, validating it, and then applying it to our data seemed an excessively difficult task to accomplish for this project.

Instead, the qualitative method of thematic analysis was applied to the process, with the raw material being the initial cardiologist's interpretations and the text version of the classifier output. This method of thematic analysis strongly relied on the work of Richard Boyatzis. (Boyatzis, 1998) This process is ongoing at the time of the creation of this master's capstone summary. However, this document will discuss the planned stages within the thematic analysis. In brief, the goal was to determine in what ways the classifier failed to capture the meaning of the text within the cardiologist's ECG interpretations.

Familiarization with the data. Categorizing 3,500 lines of ECG interpretations into the AHA categories qualified as a deep (in fact, nearly comprehensive) dive into the data. The only "meaning" in ECGs not captured by these lines individually are the instances when one clinical idea or concept spanned multiple lines of text or occasions where one line of text altered the meaning of a subsequent line of text. In review of ECGs, this occurred exceedingly rarely.

Generating initial codes. If a line of original text could not be classified into AHA categories without complete preservation of meaning, the line will be noted as lacking complete clarity. The list of all the lines so marked should contain the vast majority of all of the "uncategorized" concepts or ideas within the ECG readings.

Searching for themes among codes. One theme might be the lack of localizing terms. Although the AHA classification system has a catchall term for "maximally" and "minimally toward a given lead" as terms 346 and 347, it doesn't have specific terms for each ECG lead like I, II, or aVL. Also, although myocardial infarctions have separate terms for certain cardiac regions (like inferior myocardial infarction as 161) there are no modifying terms for given cardiac regions. For these reasons, concepts like "inferior ischemia" or "ST-depression in II, III, and aVF" cannot be fully captured in the AHA classes without additional modification. An iterative process of examining all of the codes into various groupings based on similarity will be performed to identify themes.

Reviewing, defining and naming themes. It is important that themes are clearly delimited – it should be clearly stated what ideas a theme contains and ideas that it does not. This extra work is especially useful if another site or source of ECGs were selected to be incorporated into this process. It would be important to know, in reviewing that new data set, if there were concepts or themes not being captured and having a strong delineation of themes and what they contain would help that work.

Producing the final report. At least two actions are needed here. One is to add terms that capture the missing themes and concepts into the version of the terminology standard used for this project. This would then require reclassifying the prior text lines of the ECG reads against the newly added terms and running the classifier again. Additionally, it might be useful to have a dialog with the American Heart Association Electrocardiography and Arrhythmia Committee which designed the terminology described in Mason et al. in order to describe this project's experience with the terminology they created as well as possibilities for improvement of change.

De-identification of Study Data to Allow Public Release

There are very specific requirements for the release of de-identified medical data for public use. The limitations imposed by both federal and state laws must be satisfied. The primary regulation is the Health Insurance Portability and Accountability Act (HIPAA). It contains the Standards for Privacy of Individually Identifiable Health Information, also known as the Privacy Rule. This was initially implemented 2003. This has been amended in the final HITECH Omnibus Rule. The Privacy Rule in its current manifestation provides two methods for de-identifying PHI to allow for public release of data. The first is the use of an expert to assess the risk of re-identification for all data elements submitted for release. The second is the so-called Safe Harbor method, where two conditions have to be satisfied. Eighteen different categories of PHI have to be removed from the data and the covered entity releasing the data cannot have "actual knowledge 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.” (DHHS, 2015)

Given the complexities and liabilities involved in de-identification, guidance from the NYU Medical Center Information Security Officer (ISO) was sought. This body reviewed the plan for de-identification described below and felt that it complied with hospital policy, research ethics, state and national statutes. A final review of the de-identified material would be conducted before release as well as a plan for hosting or disseminating that material.

Appendix C contains two lists – a partial list of initial data elements within the Demographics subset of the research data and the same data elements as they would appear in the publically-released data set. These are presented together to give examples of how the public data has to be restricted (items removed or masked) to make it suitable for release. The following section will describe the process of de-identification. Working methodically through this initial list of private data, each item was assessed for whether it fell into the group of 18 classes of information that were not allowed under the Safe Harbor rule. Each case would be identified by the output of a hash function – due to the need to maintain the security of the data, the source text for the hash cannot be shared here, except to say that it is not anything related to the given patient or their data directly. A unique PatientStudyID was generated for each patient. Although not specifically required by the Privacy Rule, it was determined that we would use abstracted values to represent all of the staff members (ordering physicians, cardiologists, nurses and techs obtaining studies, etc) within the data. Additionally, any free text fields (aside from ECG readings) were eliminated. Although the Privacy Rule allows ages of patients to be represented to 89, we limited the public dataset to those with records obtained on patients from 18 to 80 years of age.

Specific handling of the date and datetime elements within the database was necessary. There are various options for de-identification of this kind of data. The easiest method would have been to preserve only the year of date or datetime elements. Unfortunately, this would result in a significant loss of meaning within the data as many time intervals between events within the process of ordering, obtaining and interpreting ECGs occur within minutes of each other; very little ability to evaluate intervals between events would be preserved if all date and datetime elements were restricted to years alone. This hurdle is usually solved by one of two methods – preserving only the time intervals between the specific time points in a given case or creating a random time offset on a case-by-case basis to adjust the date and datetime values while preserving, in a secure location, the case ID with the specific time offset used. We decided for a variation of the latter, creating a random time offset for each patient rather than each individual study. This preserved not only the time interval data within each study but also the time interval data between separate ECGs performed on the same patient in the data. The exact means of how this process was designed cannot be revealed as this could be considered revealing the method used to de-identify PHI elements and thus be in violation of the second part of the Safe Harbor rule requiring methods of de-identification to remain secret.

Dates within the text interpretation of the ECGs were already removed by the process described above in the Methods section. Any free text fields (including the lines in ECG readings) were additionally searched using regular expressions for presence of dates and numbers (specifically focusing on numbers formatted like NYU medical record numbers or other identifying values like Social Security Numbers, etc.). There were very few occurrences of stray dates and no use of other identifying numbers. The few dates were redacted manually in the processed JSON files. The text in the readings was additionally searched against a list of the most common first and last names in the United States. Except for the expected matches on named clinical entities (e.g. Wolff-Parkinson-White) no matches were found.

Although it is beyond the scope of this work, it is worth having a brief discussion on the application of de-identification processes on the ECG waveform data. Are the waveform tracings of ECGs identifying? Clearly, if person A has an ECG with rare findings and you were presented with 20 ECGs of patients which included a separate ECG of patient A, you could probably re-identify Patient A from the comparison of the tests, due to the rarity of the findings in Patient A's ECG. This mental exercise assumes, however, that you possess the identified ECGs of the 20 patients, and as such, you would already have a disclosed version of all of the information (the findings in a given patient's ECG) that you were trying to keep de-identified. It is unlikely that a patient could be identified from an ECG waveform without having some prior knowledge of the appearance of a given patient's ECG tracings. Given the number of ECGs in this study, even rare findings appear in modest numbers, which makes associating a specific patient to a given pattern of findings in an ECG fairly difficult. It has been shown that mere removal of dates and times from panels of laboratory data may not effectively de-identify the data. Researchers could match a patient with an initially-identified lab to be inferred from the pattern of lab values alone. (Cimino, 2012) It is some reassurance, however, that this work required the agent identifying cases to have in their possession the original, identified copy of results of a patient to which comparisons were made.

There are methods that researchers have put forward for altering the ECG tracing in an effort to de-identify it – adjusting each time-series value (the amount of millivolts measured at a given time point) in the waveform by small amounts, amounts which would not alter the ECG interpretation but would render the electronic version of the file distinct from the original.

Future Work

There is significant opportunity to improve or expand upon the work already accomplished.

1. Cleaning the text in reports – Some of the cardiology-entered text has spelling mistakes and other minor errors. Although initially left in during the processing and classification process, these can be identified and corrected to present a more polished dataset.
2. Identifying cases with currently unassigned classification terms – Part of the hope of the classification work was that in this eventual set of over 80,0000 patients, a

few cases which fall into each class within the terminology would be identified. It would be possible to use the clinical data in the electronic health record to identify patients with conditions that might have findings not discovered by the machine or cardiologist's interpretation of their ECG. This might allow a larger scope of clinical findings to be captured in the dataset.

3. Adjudicating confusing findings – There were certain findings in the ECG reads which were impossible to correctly parse into one of two (or more) similar terms. For instance, ECG readings containing the term "atrial tachycardia" were often not specified in the text of the reading as being unifocal or multifocal in origin, and these are separate terms (52 and 53 respectively) in the terminology. It will be possible to manually review these ECGs and correct the term based on the repeat interpretation of the original ECG.
4. Restructuring the classification ontology – Currently, an ECG finding like sinus rhythm is the same type of object as a modifier like "recent" or even contractions like "and". There is the opportunity to revise the classifier so that nouns (like a specific finding) can be related to modifiers in a tree-like structure, more akin to diagramming sentences. "Sinus rhythm with frequent premature ventricular contractions" can be represented in the data not as the four items "sinus rhythm," "with," "frequent" and "premature ventricular contractions" but rather in a data structure which preserves that "frequent" is modifying "premature ventricular contractions" and that these are travelling "with" the "sinus rhythm". Structured this way, the ontology would help preserve a greater depth of meaning within the classifier. As an example, it would be difficult to search in the current data

structure for the concept of “possible ischemia”; you could search for ECGs with the term “possible” and the term “ischemia” but you might pick up in your search ECG where “possible” referred to other things than “ischemia”. If the ontology preserved that “possible” modified “ischemia”, this search would be easy.

5. Adding additional sources of ECG data – The structure of and methodology behind this dataset would be expandable to other sources of data. For starters, NYU Langone performs ECGs in many more clinical areas than the emergency department. Additionally, NYU itself has many more hospitals than Langone, all using the MUSE ECG system. More broadly, GE ECG products lead the market in the United States – exact details of market share were difficult to determine, as most market research companies offer this information only in rather expensive paid reports.* The Veterans Health Administration is just one example of an external healthcare provider that exclusively uses the GE MUSE system.

   Data from additional clinical sites could be incorporated into the current database, if the database structure was changed slightly. The StudyIDs for patients and physicians could be organized so that they are unique to a location. For instance, the StudyIDs for NYU Langone patients all begin with “100” followed by 6 additional digits to identify the patient. Another hospital system could be given the numbers starting at “200” or “200”. The ID hash that defines a given case is unlikely to overlap with prior studies at other sites (unless cases into the many millions or billions are accumulated) and overlap could be checked on creation to prevent duplicates at the time of merger with the database. If it were needed, a strategy for matching patients who received treatment at more than one clinical site would have to be developed.

6. Using the database as a supervised learning set – This database will eventually contain both findings present within the ECG captured in a terminology and the waveform data of those ECGs. It might be possible to use this database as a supervised learning set for a neural network or other algorithmic learning system; the system could be trained using the classified terms to identify features within the waveform tracings. These would be akin to supervised learning systems which are trained to identify people through facial recognition or identify pictures of cats. Historically, a waveform was normalized and parsed heuristically (the QRS complexes identified, the deviations belonging to P waves identified, intervals between waves measured, etc). The system would deduce the findings from interpretation of these mathematical transformations of the waveform. It is possible to imagine that a supervised learning system could correctly identify features without this deconstruction of the tracing.

Summary and Conclusions

A large, publicly-available, de-identified database of ECG data that includes waveforms and ECG findings classified against a terminology standard does not exist at present by search of publications or Internet sources. This work creates such a resource. The method applied to this work benefitted strongly from fairly structured entry of the ECG interpretations. This occurred due to dedicated use of the MUSE Cardiology system and also the decision to restrict the ECGs from this project to one hospital and thus a limited set of cardiologists performing readings. This method is applicable to any system of computerized ECG records, however, the less structured the data (for example, the more free text entered into ECG readings or other fields) the more work it will take to apply this method to another source of data. The team involved in this project looks forward to the public release of this dataset and is eager to see the various ways in which this data is eventually used.

References

Bousseljot R, K. D., Schnabel A. (1995). Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. *Biomedizinische Technik, 40*(1), S 317.

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*: Sage Publications.

Brown, B. D., Badilini, F. (2005). HL7 aECG Implementation Guide. Retrieved from https://www.hl7.org/documentcenter/public_temp_6C234367-1C23-BA17-0CCDE6BD4C5D9758/wg/rcrim/annecg/aECG%20Implementation%20Guide%202005-03-21%20final%203.pdf

Chazard, E., Marcolino, M. S., Dumesnil, C., Caron, A., Palhares, D. M., Ficheur, G., . . . Ribeiro, A. L. (2015). One Million Electrocardiograms of Primary Care Patients: A Descriptive Analysis. *Stud Health Technol Inform, 216*, 69-73.

Cimino, J. J. (2012). The false security of blind dates: chrononymization's lack of impact on data privacy of laboratory data. *Appl Clin Inform, 3*(4), 392-403. doi:10.4338/ACI-2012-07-RA-0028

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull, 70*(4), 213-220.

Couderc, J. P. (2010). A unique digital electrocardiographic repository for the development of quantitative electrocardiography and cardiac safety: the Telemetric and Holter ECG Warehouse (THEW). *J Electrocardiol, 43*(6), 595-600. doi:10.1016/j.jelectrocard.2010.07.015

DHHS. (2015, Nov 6, 2015). Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Retrieved from https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., . . . Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation, 101*(23), E215-220.

Jager, F., Taddei, A., Moody, G. B., Emdin, M., Antolic, G., Dorn, R., . . . Mark, R. G. (2003). Long-term ST database: a reference for the development and evaluation of automated ischaemia detectors and for the study of the dynamics of myocardial ischaemia. *Med Biol Eng Comput, 41*(2), 172-182.

Kadish, A. H., Buxton, A. E., Kennedy, H. L., Knight, B. P., Mason, J. W., Schuger, C. D., . . . Weitz, H. H. (2001). ACC/AHA clinical competence statement on electrocardiography and ambulatory electrocardiography. A report of the ACC/AHA/ACP-ASIM Task Force on Clinical Competence (ACC/AHA Committee to Develop a Clinical Competence Statement on Electrocardiography and Ambulatory Electrocardiography). *J Am Coll Cardiol, 38*(7), 2091-2100.

Kim, Y. G., Shin, D., Park, M. Y., Lee, S., Jeon, M. S., Yoon, D., & Park, R. W. (2017). ECG-ViEW II, a freely accessible electrocardiogram database. *PLoS One, 12*(4), e0176222. doi:10.1371/journal.pone.0176222

Ledezma, C. A., Severeyn, E., Perpiñán, G., Altuve, M., & Wong, S. (2014). A new on-line electrocardiographic records database and computer routines for data analysis. *Conf Proc IEEE Eng Med Biol Soc, 2014*, 2738-2741. doi:10.1109/EMBC.2014.6944189

Mason, J. W., Hancock, E. W., Gettes, L. S., Bailey, J. J., Childers, R., Deal, B. J., . . . Heart Rhythm, S. (2007). Recommendations for the standardization and interpretation of the electrocardiogram: part II: electrocardiography diagnostic statement list a scientific statement from the American Heart Association Electrocardiography and Arrhythmias Committee, Council on Clinical Cardiology; the American College of Cardiology Foundation; and the Heart Rhythm Society Endorsed by the International Society for Computerized Electrocardiology. *J Am Coll Cardiol, 49*(10), 1128-1135. doi:10.1016/j.jacc.2007.01.025

Moody, G., Muldrow WE, Mark RG. (1984). A noise stress test for arrhythmia detectors. . *Computers in Cardiology, 11*, 381-384. doi:doi:10.13026/C2HS3T

Moody, G. B., & Mark, R. G. (2001). The impact of the MIT-BIH arrhythmia database. *IEEE Eng Med Biol Mag, 20*(3), 45-50.

Popa, T. R., Mocanu, A.C. (2011). Medical Data Storage, Visualization and Interpretation: A Case Study Using a Proprietary ECG XML Format. *Annals of the University of Craiova Series: Automation, Computers, Electronics and Mechatronics, 8*(36), 44-49.

Pusic, M., Pecaric, M., & Boutis, K. (2011). How much practice is enough? Using learning curves to assess the deliberate practice of radiograph interpretation. *Acad Med, 86*(6), 731-736. doi:10.1097/ACM.0b013e3182178c3c

Pusic, M. V., Andrews, J. S., Kessler, D. O., Teng, D. C., Pecaric, M. R., Ruzal-Shapiro, C., & Boutis, K. (2012). Prevalence of abnormal cases in an image bank affects the learning of radiograph interpretation. *Med Educ, 46*(3), 289-298. doi:10.1111/j.1365-2923.2011.04165.x

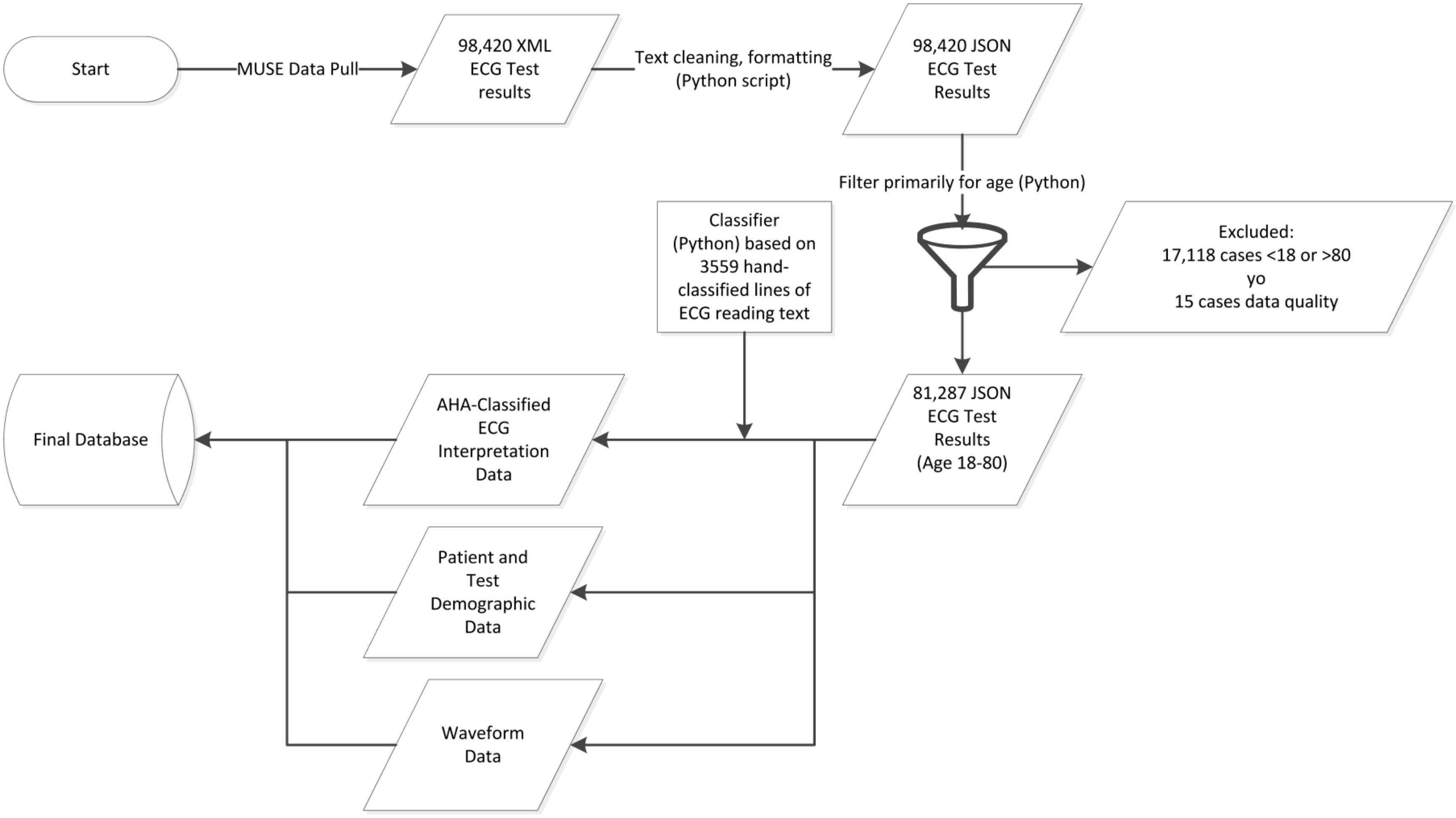
Taddei, A., Distante, G., Emdin, M., Pisani, P., Moody, G. B., Zeelenberg, C., & Marchesi, C. (1992). The European ST-T database: standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiography. *Eur Heart J, 13*(9), 1164-1172.

Appendix A – References to publically available ECG datasets

Source/Reference	Contents	Public
(Chazard et al., 2015)	1.9 million ECGs interpreted by Telehealth Network of Minas Gerais, Brazil	No
(Couderc, 2010)	34 subjects, thorough QT study, 24h 3-lead Holter 70 subjects, thorough QT study, 24h 12-lead Holter 271 subjects with coronary artery disease, 24h 3-lead Holter 93 subjects with acute myocardial infarction, 24h 3-lead Holter 201 normal subjects, 24 hour 3-lead Holter 6 subjects with torsade de pointes, 12 hour 12-lead Holter 34 subjects with history of torsade de pointes, 20 minute 12-lead ECG 73 subjects with atrial fibrillation, 10 minute 12-lead ECG	Yes Yes Yes Yes Yes Yes Yes Yes
(Kim et al., 2017)	ECG-View II, contains data from 979,273 ECGs from 461,178 subjects. Does not have waveforms of ECGs, just interval measurements. However, does have linked patient data including medications and medical conditions	Yes
(Ledezma, Severeyn, Perpiñán, Altuve, & Wong, 2014)	72 subjects with ischemic cardiomyopathy, 3-lead ECG 20 subjects with ischemic preconditioning, 3-lead ECG 51 subjects with diabetes, 8-lead ECG 25 subjects with metabolic syndrome, 12-lead ECG	Yes Yes Yes Yes
(Goldberger et al., 2000)	PhysioNet, an online resource which houses the following **incomplete** list of databases	
	European ST-T Database – 79 subjects with myocardial ischemia, 90 annotated recordings in 2 leads (Taddei et al., 1992)	Yes
	Long-Term ST Database – 86 annotated records of 2 or 3 leads each (Jager et al., 2003)	Yes
	MIH-BIH Arrhythmia Database – 47 subjects, 48 annotated half-hour excerpts of 2-lead data (G. B. Moody & Mark, 2001)	Yes
	MIH-BIH Noise Stress Database – total of 15 half-hour recordings of ECGs with various levels of signal noise (G. Moody, Muldrow WE, Mark RG, 1984)	Yes

	PTB Diagnostic ECG Database – 290 subject, 590 recordings with various clinical conditions recorded with 15 leads (Bousseljot R, 1995)	Yes

Appendix B – Data flow diagram for project


Start
MUSE Data Pull
98,420 XML ECG Test results
Text cleaning, formatting (Python script)
98,420 JSON ECG Test Results
Filter primarily for age (Python)
Excluded:
17,118 cases <18 or >80 yo
15 cases data quality
Classifier (Python) based on 3559 hand-classified lines of ECG reading text
81,287 JSON ECG Test Results (Age 18-80)
AHA-Classified ECG Interpretation Data
Patient and Test Demographic Data
Waveform Data
Final Database

Appendix C – Examples of output of processing stages of data files

Partial extract of original ECG XML from MUSE database

```
  <Diagnosis>
    <Modality>RESTING</Modality>
    <DiagnosisStatement>
      <StmtText>Atrial flutter</StmtText>
    </DiagnosisStatement>
    <DiagnosisStatement>
      <StmtFlag>ENDSLINE</StmtFlag>
      <StmtText>with variable A-V block</StmtText>
    </DiagnosisStatement>
    <DiagnosisStatement>
      <StmtFlag>ENDSLINE</StmtFlag>
      <StmtText>Abnormal ECG</StmtText>
    </DiagnosisStatement>
    <DiagnosisStatement>
      <StmtText>When compared with ECG of</StmtText>
    </DiagnosisStatement>
    <DiagnosisStatement>
      <StmtFlag>ENDSLINE</StmtFlag>
      <StmtText>XX-DEC-XXXX XX:XX,</StmtText>
    </DiagnosisStatement>
    <DiagnosisStatement>
      <StmtText>Atrial flutter</StmtText>
    </DiagnosisStatement>
    <DiagnosisStatement>
      <StmtText>has replaced</StmtText>
    </DiagnosisStatement>
    <DiagnosisStatement>
      <StmtFlag>ENDSLINE</StmtFlag>
      <StmtText>Sinus rhythm</StmtText>
    </DiagnosisStatement>
  </Diagnosis>
```

Partial extract of XML after processing into JSON format

```
"DiagnosisLines": [
          {
            "LineText": "Atrial flutter with variable A-V block",
            "OriginalText": "Atrial flutter with variable A-V block",
            "LineUserinsert": "False",
            "ComparisonFlag": "False"
          },
          {
            "LineText": "Abnormal ECG",
            "OriginalText": "Abnormal ECG",
            "LineUserinsert": "False",
            "ComparisonFlag": "False"
          },
          {
            "LineText": "When compared with ECG of a month ago,",
            "OriginalText": "When compared with ECG of XX-DEC-XXXX XX:XX,",
            "LineUserinsert": "False",
            "ComparisonFlag": "True"
          },
          {
            "LineText": "Atrial flutter has replaced Sinus rhythm",
            "OriginalText": "Atrial flutter has replaced Sinus rhythm",
            "LineUserinsert": "False",
            "ComparisonFlag": "True"
          }
        ]
```

Extract of ECG reading after categorization into the terminology standard

```
Diagnosis": {
        "CategoriesDiagnosis": "51,319,86,3,-1,-1"
        "CurrentFindingsDiagnosis": "-1,-1,51"
       "PriorFindingsDiagnosis": "-1,-1,20",
        "OriginalDiagnosisLinesCategorized": [
          {
            "FullyCategorized": "True",
            "LineText": "atrial flutter with variable a-v block",
            "OriginalText": "atrial flutter with variable a-v block",
            "CurrentFindings": "-1",
            "PriorFindings": "-1",
            "LineUserinsert": "False",
            "ComparisonFlag": "False",
            "Categories": "51,319,86"
          },
          {
            "FullyCategorized": "True",
            "LineText": "abnormal ecg",
            "OriginalText": "abnormal ecg",
            "CurrentFindings": "-1",
            "PriorFindings": "-1",
            "LineUserinsert": "False",
            "ComparisonFlag": "False",
    "Categories": "3"
},
{
    "FullyCategorized": "True",
    "LineText": "when compared with ecg of a month ago,",
    "OriginalText": "when compared with ecg of XX-dec-XXXX XX:XX,",
    "CurrentFindings": "",
    "PriorFindings": "",
    "LineUserinsert": "False",
    "ComparisonFlag": "True",
    "Categories": "-1"
},
{
    "FullyCategorized": "True",
    "LineText": "atrial flutter has replaced sinus rhythm",
    "OriginalText": "atrial flutter has replaced sinus rhythm",
    "CurrentFindings": "51",
    "PriorFindings": "20",
    "LineUserinsert": "False",
    "ComparisonFlag": "True",
    "Categories": "-1"
},
```

Appendix D - Incomplete list of AHA classifier categories

ID	Name	Category	Type
1	Normal ECG	Primary	Overall Interpretation
2	Otherwise normal ECG	Primary	Overall Interpretation
3	Abnormal ECG	Primary	Overall Interpretation
4	Uninterpretable ECG	Primary	Overall Interpretation
10	Extremity electrode reversal	Primary	Technical conditions
11	Misplaced precordial electrode(s)	Primary	Technical conditions
12	Missing lead(s)	Primary	Technical conditions
13	Right-sided precordial electrode(s)	Primary	Technical conditions
20	Sinus rhythm	Primary	Sinus node rhythms and arrhythmias
21	Sinus tachycardia	Primary	Sinus node rhythms and arrhythmias
22	Sinus bradycardia	Primary	Sinus node rhythms and arrhythmias
23	Sinus arrhythmia	Primary	Sinus node rhythms and arrhythmias
24	Sinoatrial block, type I	Primary	Sinus node rhythms and arrhythmias
30	Atrial premature complex(es)	Primary	Supraventricular arrhythmias
31	Atrial premature complexes, nonconducted	Primary	Supraventricular arrhythmias
32	Retrograde atrial activation	Primary	Supraventricular arrhythmias
33	Wandering atrial pacemaker	Primary	Supraventricular arrhythmias
34	Ectopic atrial rhythm	Primary	Supraventricular arrhythmias
35	Ectopic atrial rhythm, multifocal	Primary	Supraventricular arrhythmias
50	Atrial fibrillation	Primary	Supraventricular tachyarrhythmias
51	Atrial flutter	Primary	Supraventricular tachyarrhythmias
52	Ectopic atrial tachycardia, unifocal	Primary	Supraventricular tachyarrhythmias
60	Ventricular premature complex(es)	Primary	Ventricular arrhythmias
61	Fusion complex(es)	Primary	Ventricular arrhythmias
70	Ventricular tachycardia	Primary	Ventricular tachyarrhythmias
71	Ventricular tachycardia, unsustained	Primary	Ventricular tachyarrhythmias
72	Ventricular tachycardia, polymorphous	Primary	Ventricular tachyarrhythmias
73	Ventricular tachycardia, torsades de pointes	Primary	Ventricular tachyarrhythmias
74	Ventricular fibrillation	Primary	Ventricular tachyarrhythmias
75	Fascicular tachycardia	Primary	Ventricular tachyarrhythmias
76	Wide-QRS tachycardia	Primary	Ventricular tachyarrhythmias
80	Short PR interval	Primary	Atrioventricular conduction
81	AV conduction ratio N:D	Primary	Atrioventricular conduction
82	Prolonged PR interval	Primary	Atrioventricular conduction
100	Aberrant conduction of supraventricular beat(s)	Primary	Intraventricular and intra-atrial conduction
101	Left anterior fascicular block	Primary	Intraventricular and intra-atrial conduction
102	Left posterior fascicular block	Primary	Intraventricular and intra-atrial conduction
104	Left bundle-branch block	Primary	Intraventricular and intra-atrial conduction
105	Incomplete right bundle-branch block	Primary	Intraventricular and intra-atrial conduction
106	Right bundle-branch block	Primary	Intraventricular and intra-atrial conduction
107	Intraventricular conduction delay	Primary	Intraventricular and intra-atrial conduction
120	Right-axis deviation	Primary	Axis and voltage
121	Left-axis deviation	Primary	Axis and voltage
122	Right superior axis	Primary	Axis and voltage
123	Indeterminate axis	Primary	Axis and voltage
140	Left atrial enlargement	Primary	Chamber hypertrophy or enlargement

141	Right atrial enlargement	Primary	Chamber hypertrophy or enlargement
142	Left ventricular hypertrophy	Primary	Chamber hypertrophy or enlargement
143	Right ventricular hypertrophy	Primary	Chamber hypertrophy or enlargement
144	Biventricular hypertrophy	Primary	Chamber hypertrophy or enlargement
145	ST deviation	Primary	ST segment, T wave, and U wave
146	ST deviation with T-wave change	Primary	ST segment, T wave, and U wave
147	T-wave abnormality	Primary	ST segment, T wave, and U wave
148	Prolonged QT interval	Primary	ST segment, T wave, and U wave
160	Anterior MI	Primary	Myocardial infarction
161	Inferior MI	Primary	Myocardial infarction
162	Posterior MI	Primary	Myocardial infarction
163	Lateral MI	Primary	Myocardial infarction
165	Anteroseptal MI	Primary	Myocardial infarction
166	Extensive anterior MI	Primary	Myocardial infarction
173	MI in presence of left bundle-branch block	Primary	Myocardial infarction
174	Right ventricular MI	Primary	Myocardial infarction
180	Atrial-paced complex(es) or rhythm	Primary	Pacemaker
181	Ventricular-paced complex(es) or rhythm	Primary	Pacemaker
182	Ventricular pacing of non–right ventricular apical origin	Primary	Pacemaker
183	Atrial-sensed ventricular-paced complex(es) or rhythm	Primary	Pacemaker
184	AV dual-paced complex(es) or rhythm	Primary	Pacemaker
200	Acute pericarditis	Suggests	
201	Acute pulmonary embolism	Suggests	
202	Brugada abnormality	Suggests	
203	Chronic pulmonary disease	Suggests	
204	CNS disease	Suggests	
205	Digitalis effect	Suggests	
206	Digitalis toxicity	Suggests	
220	Acute ischemia	Consider	
221	AV nodal reentry	Consider	
222	AV reentry	Consider	
301	Borderline	Modifier, general	
302	Consider	Modifier, general conjunction	
303	Increased	Modifier, general	
304	Intermittent	Modifier, general	
311	Or	Modifier, general conjunction	
312	Possible	Modifier, general	
319	With	Modifier, general conjunction	
320	And	Modifier, general conjunction	
321	Nonspecific	Modifier, general	
322	Versus	Modifier, general conjunction	
330	Acute	Modifier, myocardial infarction	
331	Recent	Modifier, myocardial infarction	
332	Old	Modifier, myocardial infarction	
333	Of indeterminate age	Modifier, myocardial infarction	
334	Evolving	Modifier, myocardial infarction	
340	Couplets	Modifier, arrhythmias and tachyarrhythmias	
341	In a bigeminal pattern	Modifier, arrhythmias and tachyarrhythmias	
342	In a trigeminal pattern	Modifier, arrhythmias and tachyarrhythmias	
343	Monomorphic	Modifier, arrhythmias and tachyarrhythmias	
360	?0.1 mV	Modifier, repolarization abnormalities	
361	?0.2 mV	Modifier, repolarization abnormalities	

362	Depression	Modifier, repolarization abnormalities	
363	Elevation	Modifier, repolarization abnormalities	
400	No significant change	Comparison statements	
401	Significant change in rhythm	Comparison statements	
402	New or worsened ischemia or infarction	Comparison statements	
403	New conduction abnormality	Comparison statements	
501	Abnormal qrs-t angle, consider primary t wave abnormality	Primary	ST segment, T wave, and U wave
502	Normal axis	Primary	Axis and voltage
503	Repolarization abnormality	Primary	ST segment, T wave and U wave
407	Significant changes have occurred	Comparison	Comparison statements
504	Late transition	Primary	Axis and voltage
505	Wolff-Parkinson-White	Secondary	Suggests

Source of items (Kadish et al., 2001)

Appendix E - Comparison of public- and private-facing data after de-identification

Incomplete list of data elements within the private-facing research database

IDHash	Unique hash value of the given ECG case
File	File location of ECG on NYU computer system
PatientStudyID	Study-specific unique ID for patient
PatientID	Patient identification number in NYU EHR system
PatientLastName	Patient’s last name
PatientFirstName	Patient’s first name
DateofBirth	Patient’s date of birth
PatientAge	Patient’s age
StudySecondsOffset	Number of seconds subtracted from all times in patient record to mask patient data in public-facing database
Gender	Patient’s gender
Race	Patient’s race
AcquisitionDateTime	Datetime ECG was acquired
OverreaderStudyID	Study-specific unique ID for cardiology who overread ECG
OverreaderID	Cardiologist’s unique ID number in the NYU EHR
OverreaderLastName	Cardiologist overreader’s last name
OverreaderFirstName	Cardiologist overreader’s first name
OrderingMDstudyID	Study-specific unique ID for ordering physician
OrderingMDHISID	Ordering physician of ECG’s unique ID number in the NYU EHR
OrderingMDLastName	Ordering physician’s last name
OrderingMDFirstName	Ordering physician’s first name
…	

Incomplete list of data elements within the public-facing database

IDHash	Unique hash value of the given ECG case
PatientStudyID	Study-specific unique ID for patient
PatientAge	Patient's age; all patients in study are less than 90 years old
Gender	Patient's gender
Race	Patient's race
AcquisitionDateTimeAdj	Datetime ECG was acquired adjusted by privately held time offset
OverreaderStudyID	Study-specific unique ID for cardiology who overread ECG
OrderingMDstudyID	Study-specific unique ID for ordering physician
…	