# Evaluation of patient re-identification using laboratory test orders and mitigation via latent space variables

Kipp W. Johnson[1*], Jessica K. De Freitas[1*], Benjamin S. Glicksberg[2], Jason R. Bobe[1], Joel T. Dudley[1#]

[1]*Institute for Next Generation Healthcare*
*Department of Genetics and Genomics Sciences,*
*Icahn School of Medicine at Mount Sinai,*
*770 Lexington Ave 15th Fl.*
*New York, NY 10065, USA*

[2]*Bakar Computational Health Sciences Institute*
*The University of California San Francisco*
*San Francisco, CA 10065, USA*

**Authors contributed equally*
[#]*Corresponding author: joel.dudley@mssm.edu*

Anonymized electronic health records (EHR) are often used for biomedical research. One persistent concern with this type of research is the risk for re-identification of patients from their purportedly anonymized data. Here, we use the EHR of 731,850 de-identified patients to demonstrate that the average patient is unique from all others 98.4% of the time simply by examining what laboratory tests have been ordered for them. By the time a patient has visited the hospital on two separate days, they are unique in 72.3% of cases. We further present a computational study to identify how accurately the records from a single day of care can be used to re-identify patients from a set of 99 other patients. We show that, given a single visit's laboratory orders (even without result values) for a patient, we can re-identify the patient at least 25% of the time. Furthermore, we can place this patient among the top 10 most similar patients 47% of the time. Finally, we present a proof-of-concept technique using a variational autoencoder to encode laboratory results into a lower-dimensional latent space. We demonstrate that releasing latent-space encoded laboratory orders significantly improves privacy compared to releasing raw laboratory orders (<5% re-identification), while preserving information contained within the laboratory orders (AUC of >0.9 for recreating encoded values). Our findings have potential consequences for the public release of anonymized laboratory tests to the biomedical research community. We note that our findings do not imply that laboratory tests alone are personally identifiable. In the attack scenario presented here, reidentification would require a threat actor to possess an external source of laboratory values which are linked to personal identifiers at the start.

*Keywords:* Electronic health records, anonymization, patient re-identification, data privacy, variational autoencoder

## 1. Introduction

Electronic Health Records (EHRs) have been widely adopted as a component of the modern American healthcare system (1). EHRs contain information such as disease-related diagnosis billing codes, lab test orders and results, procedures performed, and medications prescribed. Although EHRs are primarily designed for the purpose of encounter documentation and billing, the data can also be repurposed for efforts to improved clinical care (2, 3) or for biomedical investigation (4–6).

For use in research, EHRs are often de-identified in accordance with the Health Insurance Portability and Accountability Act (HIPAA) (7). HIPAA's Privacy Rule mandates protection for identifiable variables such as name, zip code, date of birth, etc. Because of this, public release of EHR data requires either (1) expert "determination" or (2) "safe harbor" privacy practices. Expert determination involves an individual with appropriate knowledge and experience determining that data poses minimal risk. "Safe harbor" practice is the removal of 18 pieces of information from the EHR, with the 18th being a "catch-all" category for "any other unique identifying characteristic." However, the definition for what constitutes individually identifiable information has been challenged by a variety of re-identification attacks and privacy breaches (8, 9). In practice, the privacy rule does not constrain the types of uses of health data once it has been de-identified by these methods, although covered entities sometimes take additional precautions such as data use agreements that forbid intentional re-identification.

Re-identification is the process of matching anonymized personal data with its owner via linkage with an external resource. Information such as a person's name and address are obviously identifying, but in some circumstances data such as disease diagnoses or lab tests may be identifiable. In fact, there have been several important examples of this type of privacy attack. Loukides et al. demonstrated that existing privacy protection methods were not sufficient to protect against re-identification by identifying a subset of 2800 patients from using EHR diagnosis codes alone (10). Although the diagnosis code dataset from EHRs were anonymized, the risk for re-identification came from cross-referencing with a secondary data source that contained the patient's exact diagnosis codes. Other researchers have developed strategies to anonymize combinations of disease billing codes with linked demographics (11).

In this manuscript, we first demonstrate the uniqueness of the pattern of physician-ordered laboratory tests for specific individuals. After finding that these laboratory orders are highly specific, we propose an algorithm and evaluation framework to re-identify patients using only a single day of laboratory orders. Following this, we explore if latent variables can be constructed using a variational autoencoder which simultaneously preserve information contained within the laboratory orders and also increase patient privacy.

## 2. Methods

We present an overall workflow of the study in **Figure 1**. While we do use EHRs of real patients in this paper, our dataset is anonymized (i.e., de-identified) and does not include any explicit identifiers for patients such as name, social security number, hospital medical record number, or specific dates of encounters. Our dataset uses pseudo-identifiers for each patient that are internally consistent but do not map to outside datasets. All re-identification methods and results presented do not attempt to match pseudo-identifier to real identities, as that would violate ethical research practice, patient privacy, the Health Insurance Portability and Accountability Act of 1996 (HIPAA), and institutional policies of the Mount Sinai Hospital and Icahn School of Medicine at Mount Sinai.

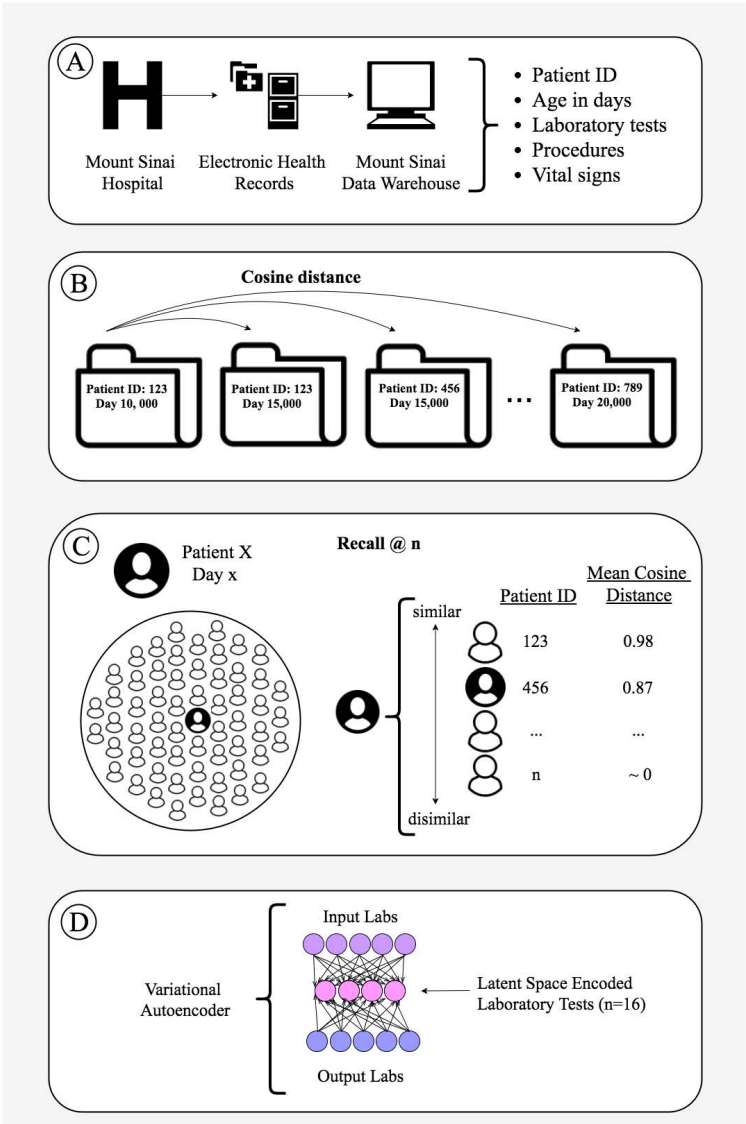


**Figure 1.** Overall workflow of our study. A) Data for this study was obtained from the Mount Sinai Hospital Data Warehouse B) Cosine distances between each patient-day event were calculated C) Evaluation by Recall @ n from n = 1 to n = 100 D) Use of a variational autoencoder to anonymize laboratory orders.

### 2.1. *Data preparation of research cohort and laboratory tests*

We used the EHRs of patient visits from the Mount Sinai Hospital (MSH), a tertiary-care urban hospital located on the Upper East Side of Manhattan in New York City. For this study, we obtained the records of all individuals between 18-90 years old. Since we sought to obtain generalized re-identifiability statistics, we did not select for patients based upon any particular criteria.

We queried the MSH EHRs for all possible laboratory tests ordered and their values. We removed laboratory tests that did not have numeric value results, could not be made to give binary data (e.g. positive/negative result), could not be used to give ordinal results (e.g. low/medium/high), had results which were clearly erroneous and nonsensical results (e.g., some labs had values which were long text strings

describing laboratory tests in place of results), or had other missing information such as order date.

## 2.2. *Assessment of how characteristic patient laboratory tests are for individual patients*

We first characterized how unique each patient's laboratory tests were. To do this, we concatenated all the laboratory tests which had been ordered at any time point for each patients into a single standardized text string. We then computed the MD5 hash of this standardized text string so that each unique combination of laboratory tests could be represented as a unique 128-bit checksum. Ultimately, patients who have received the same permutations of laboratory tests will have exactly the same MD5 hash. Finally, we checked for overlap among the MD5 hashes in order to determine the uniqueness of laboratory test orders.

## 2.3. *Assessing if using one day of patient records is sufficient to re-identify patients*

We next sought to determine if a single day of patient records would be sufficient to re-identify a patient compared to a random sample of other patients. For computational tractability we included in this analysis only those laboratory tests which had been ordered at least 500 times.

### 2.3.1. *Creation of patient-day-laboratory vectors*

Each individual patient's laboratory records were collapsed to the day in which they were ordered. If the same laboratory test was ordered more than once on the same day, we took only one occurrence of that test. We thus assembled each patient-day as a vector θ of length *l*, where *l* is the count of all laboratory tests obtained from the EHRs. Laboratory tests for a given patient-day were considered to be a binary variable where 0 denotes absence (laboratory test not ordered for this patient on this day) and 1 denotes presence (laboratory test ordered for this patient on this day)

## 2.4. *Vector distance metrics*

After computing the binary lab vector for each patient-day, we then determined pairwise similarities between patient-day vectors by computing their cosine distance. The cosine distance is a straightforward measure of similarity between vectors computed by taking the dot product of two vectors divided by the product of the two vectors' magnitude (Eq. 1).

$$\text{cosine distance} = 1 - \frac{\theta_1 \theta_2}{\|\theta_1\| \|\theta_2\|} \tag{1}$$

We thus assembled a symmetric MxM pairwise cosine distance matrix where M is the total number of patient-days. Each *(i,j)* entry in the distance matrix corresponds to the cosine distance between laboratory tests on patient-day *i* and patient-day *j*. We selected cosine distance as the similarity metric because it is a vector space metric commonly used information retrieval settings. The cosine distance in the special case of non-negative binary data (e.g. 0, 1) is also known as the Ochai distance and has a range of [0, 1] where 0 is perfect dissimilarity and 1 is perfect similarity.

### 2.5. *Patient-day re-identification algorithm*

Because the running time of pairwise distance computation grows according to the square of the patient-day counts (i.e. $O(n^2)$, quadratic complexity), it was not computationally feasible to compute the pairwise distances between all patient-day vectors. We thus posed our re-identification task as an attempt to see if, given a single-day of patient records, we could re-identify the patient based upon his or her other day's records compared to the records of 99 other randomly selected individuals.

Specifically, we randomly selected a single patient-day vector to act as the seed "breached record" for query. Our dataset for re-identification comprised of that breached patient's other patient-day vectors, not including the breached record, and all of the patient-day vectors of another 99 randomly selected patients. We then computed the cosine distance of this query vector from all other patient-day vectors in our sample. Then for each patient, we calculated the mean of the cosine distances of all their vectors from the query vector. Thus, in the end, given one patient-day record we had 100 distances corresponding to the mean distance of 100 other patients from this one patient-day. We computed this for all patient-days in the dataset. We then repeated the entire above algorithm 100 times.

For each iteration of the previous algorithm, we ultimately obtained 100 scores for distance between our query patient-record and 99 randomly selected individuals, plus the other records belonging to the initial patient from whom we extracted the seed "breached" record.

### 2.6. *Patient-day re-identification evaluation framework*

We evaluated our performance using a modified version of the "*Recall @ n*" metric commonly used in information retrieval. Since there is only one correct patient match to our query "breached" record, we evaluated if this correct patient match was within the scores corresponding to the n closest patients. The score per patient record and n was computed as a binary variable (e.g. patient is within n closest records = 1 or patient not within n closest records = 0). *Recall @ n=1* implies that the correct match was the closest score to our patient. *Recall @ n=100* will always be 100% since that implies that the patient is within the closest 100 patients queried, which will always be the case since we are querying a sample of 100 patients.

The expected *recall @ n is n/100* for a completely random classifier. Thus, we can assess our re-identification algorithm as the improvement over random classification (the null hypothesis). This formulation analogous to the area under the receiver-operating characteristic curve ("AUROC") commonly used for assessing supervised machine learning classification performance.

### 2.7. *Generalization of patient laboratory test data using a variational autoencoder*

Finally, we sought to determine whether we could encode laboratory tests orders into a reduced-dimensional latent space which was still useful but could reduce re-identifiability. To do this, we

employed a variational autoencoder implemented in Keras (https://github.com/keras-team). Variational autoencoders feature two major architectural components: First, an encoding model which takes a sequence of inputs (in our case, binary presence or absence of lab tests) and encodes them into a latent hidden representation space. A generative decoder then decodes the latent space representation back into a probability distribution representing the input data. We employed a standard VAE loss function which is the sum of the binary cross entropy between the input lab test vectors and output lab test vectors plus the Kullback-Liebler divergence between the learned encoding probability distribution and a unit Gaussian (Eq. 2).

$$l = -\sum_{x} p\left(x_{\text{out}}\right) \log q\left(x_{\text{in}}\right) + KL\left(Z(\text{latent}), N(0,1)\right) \tag{2}$$

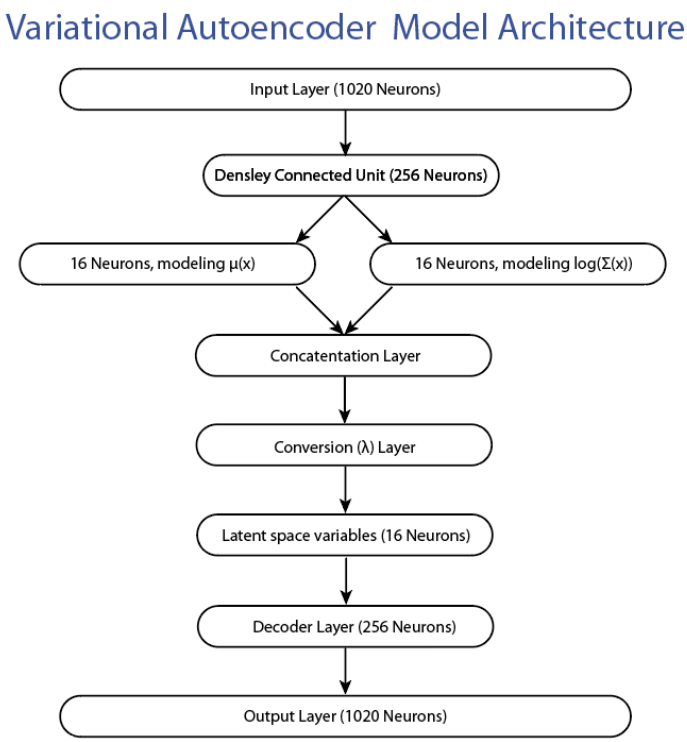


**Figure 2:** Architecture of variational autoencoder

Our hypothesis is that releasing the latent-state variables instead of the raw laboratory tests would still be useful to researchers, but would importantly reduce the potential for direct re-identifiability. After obtaining latent-space variable from this method, we employed the same technique as in the previous section (cosine distance between person-day latent variable vectors) in order to see if we could re-identify patients given one single day of data (e.g., the latent variables corresponding to that day). For ease of computation, we desired a minimal number of neurons in the latent space which could accurately recapitulate the input vectors. We found that $2^1$, $2^2$, and $2^3$ latent neurons were insufficiently accurate, but $2^4$ (16) latent space neurons produced experimentally acceptable results. The architecture of the model is given **Figure 2.**

## 3. Results

### 3.1. *Electronic health record data*

From the selected cohort of 731,850 individuals, we obtained laboratory records from those with at least one recorded laboratory test. These individuals had 342,485,583 laboratory test values for 2,635 different possible laboratory procedures. These distinct labs had been ordered on average 468.0 times per patient (standard deviation: 1,415) with a minimum frequency of one and a maximum frequency of 94,749 different laboratory results. This range of results is likely due to the fact that Mount Sinai Hospital sees a unique mix of patients, from everyday office visits to patients who may remain in the intensive care unit for weeks. The average patient had records for 49.8 different kinds of labs (standard deviation of 40.4) with a minimum of one kind of laboratory test and maximum of 442 types of different laboratory tests. The laboratory tests in total represented a period totaling 17,657 years (6,449,310 patient-days). Patients had a mean of 8.81 different days (standard deviation: 20.6) with at least one laboratory test result. The total range of days per patient was from one day to 2.47 years. 81.6% of patients had 10 or fewer days of laboratory results and 90.7% had fewer than 20 days of laboratory values. There were 218 different laboratory tests ordered only once (8.3% of all tests) and 1,186 laboratory values were ordered less than 1000 times (45.0% of all tests).

### 3.2. *What percentage of patients can be uniquely identified by laboratory tests ordered for them?*

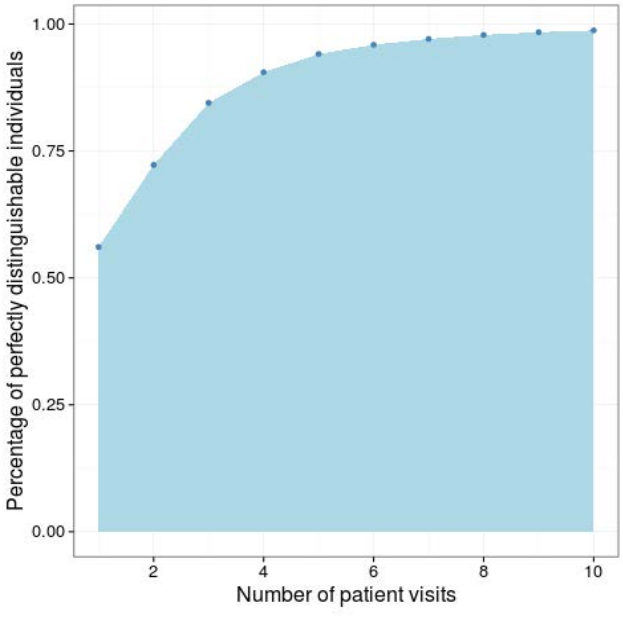


**Figure 3**: Percentage of patients whose particular pattern of visits are completely unique to them, by number of visits to hospital.

We analyzed the uniqueness of the laboratory tests ordered per patient, e.g. what percentage of patients had perfectly unique laboratory tests different from all other patients (**Figure 3**). This corresponds to the ability to perfectly recognize a patient by simply knowing what laboratory test have been ordered. We did not consider the numerical results for the lab, but merely assessed whether the tests had been ordered for a given patient or not. In total, 56.1% of patients could be perfectly characterized by their laboratory results (e.g., their particular combination of laboratory tests was completely different from all other patients in the EHR). However, the distinguishability increased very rapidly with increasing count of encounters in the EHR. Patients who had at least two days of laboratory values were different from all

other patients 72.3% of the time. Patients who had at least nine days of laboratory values (the mean number of days per patient) were different from all other patients 98.4% of the time.

### 3.3. *Can we re-identify patients using only 1 day of laboratory tests?*

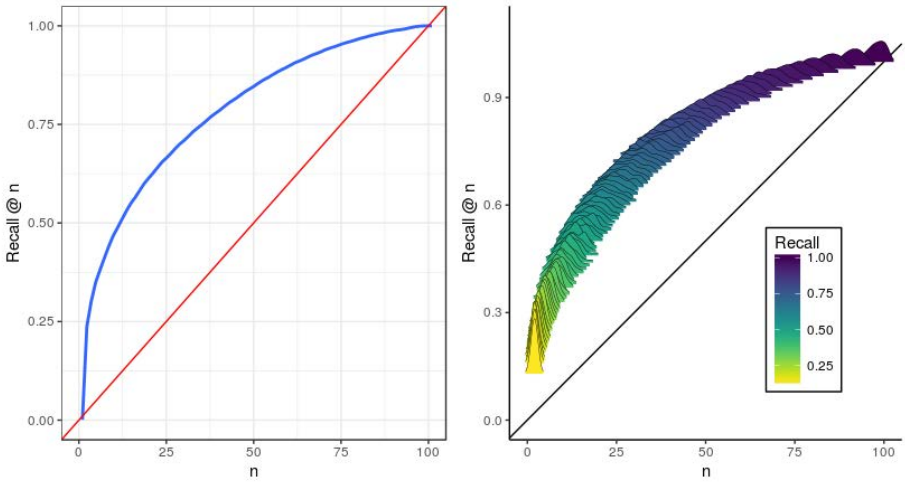


**Figure 4**: Reidentification performance using only one day of lab values. Panel on right shows distribution from 250 simulations.

We next formulated a theoretical privacy "attack": Given only a single day of records for a patient, could we re-identify this individual from a set of 99 other individuals? We show the performance for this re-identification task in **Figure 4**. Here, the red line represents the probability for random re-identification and the blue line represents the added ability to distinguish above random. We ranked the query individual as the most similar individual 25% of the time. We could place the query individual among the top 10 individuals 47% of the time.

#### 3.3.1. *Assessing the performance of latent variables from variational autoencoder to predict laboratory orders*

After applying a variational autoencoder to encode input EHR variables, we first assessed whether our encoded latent variables indeed adequately model the dataset. This is important, because we do want to ensure they retain adequate information in the data. We then attempted to predict whether a given test would be ordered for a patient or not on a given day. It is important to show that the latent variables are actually associated with laboratory results before we can demonstrate that they may be useful for anonymity.

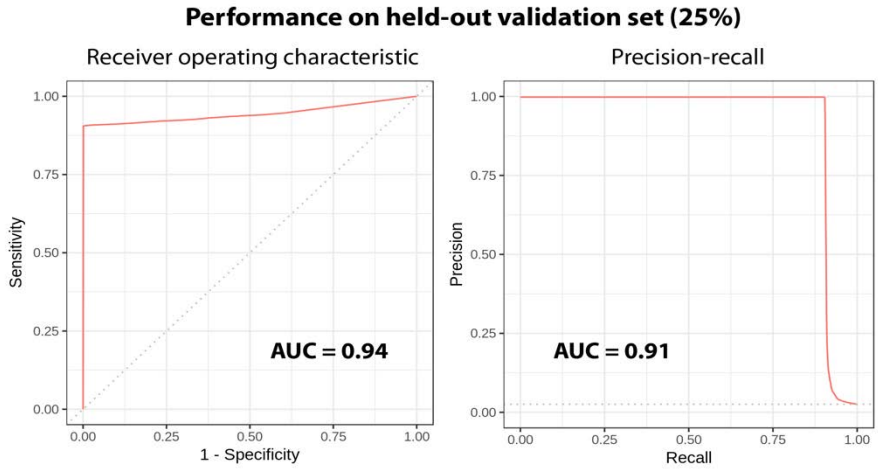


**Figure 5**: Ability of learned encodings to recapitulate input laboratory order vectors on a held-out validation set. ROC curve on left and Precision-recall curve on right.

We found that our latent variables were highly predictive of the lab order status, achieving an area under the receiver-operating characteristic curve of 0.94 and area under the precision-recall curve of 0.91, as demonstrated in **Figure 5**. This means that they recapitulate the underlying laboratory tests well.

3.3.2. *Comparing raw lab tests to latent-space abstracted laboratory tests for privacy preservation*

We assessed the ability to re-identify patients based upon cosine distance of latent variables. This is the same algorithm as used previously to re-identify patients, but with our encoded variables representing patient labs instead of using the patient labs themselves. We found that in every case, using encoded latent variables gave greater privacy protection compared to the raw lab values used for the same samples (**Figure 6**).

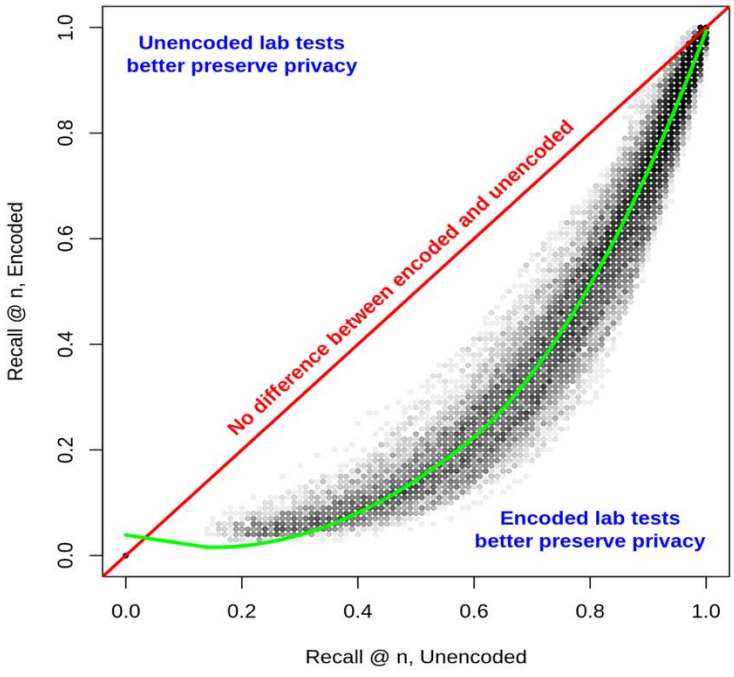


**Figure 6**: The learned latent space neurons were significantly better at reducing recall @ n for all a values of n. The red line represents the situation where the re-identification recall @ n score is the same between the unencoded lab tests and the latent space lab tests.

For *recall @ 1* through *recall @ 99* (since 100 samples were used per assessment), latent space performed significantly better at all points ($p<10^{-16}$, matched pairs t-test). We were only able to perfectly match the individual from one day of laboratory orders 5% of the time, compared to 25% of the time using the raw laboratory orders as shown previously.

## 4. Discussion

Using the Mount Sinai Hospital (MSH) patients' laboratory test history and a straightforward similarity measure, we discovered that by the time the patient has had contact with the hospital on nine separate days, their laboratory test orders are completely unique to that patient 98.4% of the time. This is a significant finding, since it implies that public datasets which contain all of the laboratory tests ordered for a specific person may be able to be matched against a known set of electronic health records (EHR) with perfect fidelity in some cases. We also show that we can obtain reasonable re-identification performance using a single day of laboratory values. Finally, we demonstrate that latent encoded variables make the problem of re-identification significantly more difficult without knowing the exact model used to encode the latent variables.

One of our primary motivations for this study stems from the idea that lab tests as are commonly used as covariates in statistical models to help to produce more accurate probability estimates for outcomes. For example, if a researcher intended to study the effect of statin therapy on incident heart disease, he or she would need to adjust for a levels of baseline LDL cholesterol and other lab tests. Instead of using actual lab tests, lower-dimensional encoded variables which contain the same amount of information as the lab tests would serve just as well as control variables. This is exactly analogous to the use of genetic SNP principal components to represent genetic ancestry in genome-wide association studies. One of the major values of our study is that we demonstrate that lower-dimensional representations of the EHR contain similar amounts of information as unprocessed records, while simultaneously preserving privacy.

Our study had several limitations. First, the work was performed with data from only one healthcare institution. However, MSH is a large tertiary care hospital with a significant diversity of patients. We also focused exclusively on laboratory test orders and did not include data such as disease diagnoses, ethnicity, gender, etc. which are often included in EHR. In our re-identification analysis, we attempted to identify an individual against a subset of 99 other random individuals, not the entire cohort of patients. Although this is a realistic scenario when performing biomedical research on a specific patient population, further analysis is needed to understand if our methods hold true when identifying an individual from the entire database. Finally, we assessed here only the binary presence or absence of laboratory test orders. It is quite possible that considering the numeric results of laboratory tests could increase re-identifiability substantially. For example, hypothetical patients with LDL cholesterol test results of 60mg/dL vs. 600mg/dL would be easily separable, although our current method considers only the fact that LDL tests were ordered for both patients. However, as we have demonstrated, considering only the binary absence or presence of orders already works reasonably well and we believe our performance metrics are conservative.

Potentially, replacing our variational autoencoder with other kinds of autoencoders or other dimensionality-reduction methods would also have been effective. Autoencoders essentially work by learning compression and decompression functions which minimize a loss function, whereas variational autoencoders learn a probability distribution which minimizes a loss function. Experimentally, our use of a VAE worked well enough to learn latent variables. By introducing this as a proof-of-concept, we felt that it would not be too valuable to benchmark against other alternatives. Future experimental and theoretical work could explore the dimensionality reduction methods used in this paper more thoroughly. Finally, we cannot release the training dataset since it contains the real patient records of hundreds of thousands of patients and could potentially enable future reidentification attacks.

We must also note here that our findings do not imply a threat model whereby patients may be identified from laboratory tests themselves, without a threat actor having an outside source of information. We show here only that lab tests are highly distinctive. For re-identification, the techniques presented here would require the threat actor to have at least some amount of information from another data source containing laboratory tests which were matched to actual patient identifiers. Furthermore, our re-identification technique only attempted to re-identify from one out of 100 instead of one out of the entire dataset, since our method for computing pairwise vector distances would not scale computationally to that extent.

Taken altogether, we believe that our findings have significant implications for the release of anonymized laboratory test results to the broad biomedical research community. Researchers should consider the possible consequences of making extensive laboratory order data for patients freely available, and should inform patients that this level of detail may potentially make them open to re-identification.

If researchers choose to release data, we suggest they consider providing latent-variable encoded laboratory values instead if this data would remain useful in their particular scientific context. Potentially, the methods we demonstrate here for laboratory test orders could be applied to other forms of data contained within the EHR.

Scientists have an obligation to respect their subjects' generosity in donation of data by maintaining their privacy and here we have demonstrated one method to make re-identification more challenging.

## References

1. Adler-Milstein J, Jha AK. HITECH Act Drove Large Gains In Hospital Electronic Health Record Adoption. *Health Aff (Millwood)* 2017;36:1416–1422.

2. Johnson KW, Torres Soto J, Glicksberg BS, et al. *Artificial Intelligence in Cardiology. J. Am. Coll. Cardiol*. 2018;71:2668–2679.

3. Johnson KW, Shameer K, Glicksberg BS, et al. Enabling Precision Cardiology Through Multiscale Biology and Systems Medicine. *JACC Basic Transl Sci* 2017;2:311–327.

4. Glicksberg BS, Johnson KW, Dudley JT. The next generation of precision medicine: observational studies, electronic health records, biobanks and continuous monitoring. *Hum. Mol. Genet*. 2018;27:R56–R62.

5. Glicksberg BS, Miotto R, Johnson KW, et al. Automated disease cohort selection using word embeddings from Electronic Health Records. *Pac Symp Biocomput* 2018;23:145–156.

6. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci Rep* 2016;6:26094.

7. Office for Civil Rights, Department of Health and Human Services. Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule and the National Instant Criminal Background Check System (NICS). *Final rule. Fed Regist* 2016;81:382–396.

8. Meeks DW, Smith MW, Taylor L, Sittig DF, Scott JM, Singh H. An analysis of electronic health record-related patient safety concerns. *J Am Med Inform Assoc* 2014;21:1053–1059.

9. Menon S, Singh H, Giardina TD, et al. Safety huddles to proactively identify and address electronic health record safety. *J Am Med Inform Assoc* 2017;24:261–267.

10. Loukides G, Denny JC, Malin B. The disclosure of diagnosis codes can breach research participants' privacy. *J Am Med Inform Assoc* 2010;17:322–327.

11. Poulis G, Loukides G, Skiadopoulos S, Gkoulalas-Divanis A. Anonymizing datasets with demographics and diagnosis codes in the presence of utility constraints. *J Biomed Inform* 2017;65:76–96.