Analysis of Clinician Evaluation of Synthetic Chronic Kidney Disease Data

Manuchehr Aminian[1], Nipuni de Silva[2], Sucharitha Dodamgodage[2], David A. Edwards[3], Darsh Gandhi[4], Nicholas Harbour[5], Hai Van Le[6], Luan Lopes[7], Adam Petrucci[8], Thabo Samakhoana[9]

[1] *Cal Poly Pomona, Pomona, CA, USA*
[2] *Clarkson University, Potsdam, NY, USA*
[3] *University of Delaware, Newark, DE, USA*
[4] *The University of Texas at Arlington, Arlington, TX, USA*
[5] *University of Nottingham, Nottingham, UK*
[6] *North Carolina State University, Raleigh, NC, USA*
[7] *Oklahoma State University, Stillwater, OK, USA*
[8] *Michigan State University, East Lansing, MI, USA*
[9] *Johns Hopkins University, Baltimore, MD, USA*

(*Communicated to* MIIR *on September ??, 2025*)

Study Group: 41st Annual Workshop on Mathematical Problems in Industry, Claremont Graduate University, Claremont, CA, June $9 - 13, 2025$.

Communicated by: Marina Chugunova

Industrial Partner: Vironix

Presenter: Zach Dana

Key Words: synthetic data, data validation, disease progression, classificator, bias model

Summary

Vironix Health Inc is an AI-driven software-as-a-service (SaaS) company focused on early detection and remote management of chronic illnesses, including chronic kidney disease (CKD). To support predictive modeling of CKD progression, Vironix has implemented machine learning models for generating synthetic electronic health records (EHRs) that supplement real-world data. Ensuring the quality of this synthetic data is essential for maintaining the fidelity of downstream predictive models.

To evaluate synthetic data quality, Vironix has employed blinded clinician assessments of anonymized patient profiles drawn from both real and synthetic datasets. However, these evaluations often exhibit large variance due to differences in clinician background and experiences. In this study, we propose a refined clinician evaluation framework that incorporates a new study design, standardized metrics (including Clinician and Patient-Based Weighting), and statistical measures such as Maximum Mean Discrepancy (MMD). We also introduce normalization methods, leveraging Generative Adversarial Networks (GAN) and Gaussian Mixture Models (GMM) applied to proxy and existing CKD datasets, to account for clinician rating variability driven by regional and experiential factors.

1 Introduction

Vironix is an industry leader in the design and implementation of AI-based products for monitoring the status and progression of chronic diseases. Their approach focuses on the collaboration between three distinct parties: individual patients, AI, and medical professionals. Through the Vironix virtual platform, patients provide health information and updates to a machine-learned program that advises both patients and their chosen physicians. Integration of these three modes in a single pipeline allows for finer data collection, more reliable and personalized medical advice, more accurate anticipation of disease progression, and improved response times when preventative measures are still available. Top-quality AI is critical to the proper functioning of the Vironix project, and by extension to the personal health of Vironix's clients and users.

Substantial quantities of data are necessary for the training of effective AI. In general settings, these requirements can surpass what is available, but the challenge is particularly acute in medicine. Established legal protections of privacy and personal data add a question of accessibility to the matter of available data volume. Adapting to these challenges, Vironix developed a process for generating synthetic data. This is used to supplement the pool of actual patient data. The combined pool is then sufficiently expansive for training AI by standard machine-learning methods.

For credibility, synthetic data should be validated by clinicians. That is, though the results of Vironix's generative algorithm perform well with respect to statistical comparison with real data, evaluation of practicing medical professionals is considered the most effective standard for AI in medicine. This is in part because humans are still believed to be the best assessors of medical condition. Given the results of their models, Vironix suspects this is especially true of unusual or outstanding clinical phenotypes. Moreover, the general public has more trust in clinical opinions than AI [23]. Consequently, clinicians' endorsement of Vironix's products will be important to their widespread adoption.

To better understand the clinical evaluation of synthetic data, Vironix appealed to MPI 2025 for a mathematical investigation. For concreteness, they proposed focusing on chronic kidney disease (CKD) and provided some survey data. The guiding inquiries:

- What factors contribute to clinicians' evaluation of data veracity?
- Given such results, how could the quality of synthetic data be improved?

2 Overview of Synthetic Data Generation

High-quality medical data is critical for healthcare-related research. Unfortunately for said research, the highly sensitive nature of medical data often precludes its public distribution. The standards for protection of such personal data fall under the regulations of the Health Insurance Portability and Accountability Act (HIPAA) in the U.S. and General Data Protection Regulation (GDPR) in the EU. Hence, researchers have designed various methods of producing quality synthetic data to compensate for the scarcity of patient data. Generally, synthetic data should have the same structural and statistical characteristics as the real data. Though there is no universally accepted evaluation met-

rics for synthetic data, there are some quantitative and qualitative metrics for assessing its quality. In terms of quantitative criteria, Naseer et al [20] proposed investigation of

(i) whether the synthetic data have the same marginal distributions as the real data,

(ii) whether the correlations in synthetic datasets resemble the real datasets,

(iii) both (i) and (ii) in combination, and

(iv) the usefulness of the synthetic data for subsequent machine learning assignments.

Multiple methods are available to satisfy these inquiries. One common tool, Maximum Mean Discrepancy (MMD) [26, 28] is a distance-based approach. MMD compares the overall similarity of two distributions by first embedding the distributions into a Hilbert space and then computing their distance in that space. Other metrics for synthetic data evaluation include α-Precision, β-Recall, and authenticity [2]. Roughly speaking, α-Precision tests whether synthetic data matches the most typical real data, while β-Recall tests whether the most typical synthetic data resembles real data; authenticity simply tests whether a synthetic model truly generates 'new' samples. Naseer et al [20] used three methods to compare real and synthetic data. The first involved analyzing the relative distributions of real and synthetic data via k-means clustering, as quantified by a log-cluster metric. The second method, so-called pairwise correlation difference (PCD), compares feature correlations within synthetic data to those within real data, as measured by the Frobenius distance. The third method was synthetic ranking agreement (SRA), which use the area under the receiver operating characteristic curve (AUROC) to compute the probability of agreement between experiments conducted with synthetic data and those conducted with real data. They also used multiple machine learning techniques, including logistic regression, random forests, and Gaussian-Naive-Bayes to compare real data and synthetic data.

When possible, blinded clinician evaluations are a common method of qualitatively assessing synthetic data. However, this method can be biased by numerous factors, including the background, experiences, and demographics of clinicians. We detail more drawbacks at the end of the next section.

Murtaza et al. [19] summarized the methods used to generate synthetic data into 3 categories: (1) knowledge-driven, (2) data-driven and (3) hybrid methods. Knowledge-driven methods such as structural methods (patient data definition language document) and behavioral methods (state transition machines like Synthea) source from public information such as human experts or academic research. Data-driven methods derive the truth from the limited real patient data available. Such methods include mixture models, copulas, Monte Carlo, and Bayesian network as well as machine learning based synthetic medical data generation such as generative adversarial networks (GANs), neural networks, autoencoders, and decision trees. Hybrid methods combine knowledge-driven and data-driven techniques, allowing researchers to understand the basic structure of real data with reinforcement from expert knowledge.

3 Chronic Kidney Disease (CKD)

For the problem presented at the workshop, Vironix asked participants to focus on chronic kidney disease. CKD affects more than 10% of the global adult population worldwide [27]. Decreased kidney function characterizes the disease, usually quantified through estimated glomerular filtration rate (eGFR) [15, 17]. CKD is classified into Stages 1–5 based on biomarkers such as eGFR, with higher eGFR values corresponding to earlier stages of the disease [31].

One important, commonly used indicator of the stage is the *kidney failure risk equation* (KFRE). There are two levels of complication: the first is the *four-variable* KFRE, which includes the following pieces of data:

- age
- gender
- estimated globular filtration rate (eGFR)
- urine albumin-creatinine ratio (UACR)

The second is an *eight-variable* KFRE which includes the data above as well as four serum compound levels.

Other biomarkers include albuminuria, serum potassium and creatinine levels, urine protein creatinine ratio (UPCR), and more [15]. There are inverse relationships between eGFR and serum creatinine as well as serum creatinine and hemoglobin, but there is a positive relationship between serum creatinine and UPCR. Diabetes and hypertension are common comorbidities and may even be causes of the disease.

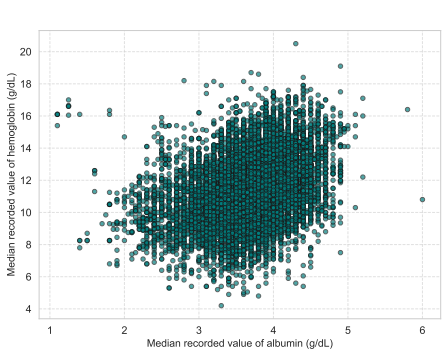
Despite biomarkers like eGFR, identification of patients with CKD is difficult and a physician's own lived experiences play an important role. In the U.S., African-American patients tend to cluster around a few physicians [3]. According to Bach *et al.* [3], the physicians who treat black patients are less likely to be board certified and report that they are less able to provide quality care to their patients than those who treat white patients. Further, minorities whose primary language is not English tend to cluster around only a few primary care physician practices [24]. To complicate matters, Vaidya and Aeddula found in a recent study [30] that CKD diagnosed by primary care physicians tends to present differently than CKD diagnosed by nephrologists.

Health inequities play a major role in the development and progression of CKD. In the U.S., African Americans are much more likely to develop CKD [14] and progress to end-stage renal disease faster than white Americans [15, 21]. Further, socioeconomic status (SES) is highly intertwined with race; low SES (determined by income, education, and occupation) corresponds to poorer prognosis for patients [22]. Low- and middle-income families may not be able to visit a nephrologist or access dialysis centers or renal transplants, and thus do not receive adequate treatment for the disease [25]. Black, Hispanic, and Asian patients were more likely to present with diabetes and severe albuminuria and were younger on average than white patients in a study conducted by Chu *et al.* [5]. Moreover, CKD has been shown to progress faster in Black and Hispanic patients [11]. Individual (such as health literacy, insurance, risk perception) and community (such as racial segregation and neighborhood poverty) SES factors influence patients' access to

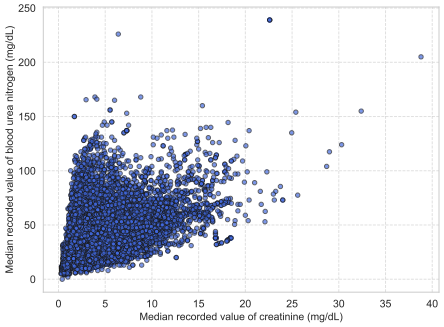
physicians, willingness to receive treatment, and quality of treatment, thus influencing CKD progression and prognosis [22]. These factors influence a physician's disposition to diagnose patients with CKD, their perceived severity of the disease, and the proposed treatment plan.

4 Chronic Kidney Disease Data from Hong et al. [10]

Research by Hong et al. [10] set the groundwork for a preliminary analysis of the important indicators of CKD. Their data on hospital admissions was used to construct a rich dataset of 29,714 CKD patients, characterized by 41 numerical and 7 categorical variables. Some of these variables were also included in patient profiles provided in the Vironix survey to practicing clinicians. These included creatinine, hemoglobin, albumin, age, and gender. Table A 1 in the appendix describes the variables related to CKD most pertinent to this project.



(a) Scatterplot of Albumin vs. Hemoglobin



(b) Scatterplot of Creatinine vs. Blood Urea Nitrogen

Figure 1. Correlations of some important variables.

This data allowed for an independent investigation of CKD indicators. Figure 1 illustrates correlations between variables that are markers of kidney function, whose relationships can reveal patterns of renal impairment. Out of 29,714 patients, 5,532 had recorded the median value of their eGFR. These were split into two equal populations, one whose median eGFR is higher than the mean of the population's median eGFR and one with eGFR lower than the mean of the population's median eGFR. The was used to train "virtual clinician" models. Based on earlier research regarding primary indicators of CKD (Section 3), the models also included variables such as albumin, creatinine, hemoglobin and blood urea nitrogen. Random Forest outperformed Logistic Regression and Neural Networks in both the high and low eGFR cases, achieving an accuracy score of over 90% (see Fig. 2.)

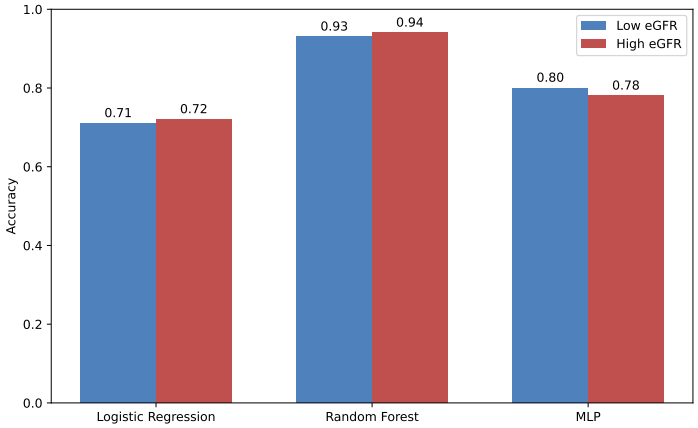


Figure 2. Classifier Accuracy by eGFR Group [10].

5 Iimori et al. [12] – Simulated Events, Synthetic Classification

The Iimori et al. [12] study concerns prognosis of CKD. To complement their study, they provided a dataset of anonymized patients longitudinal profiles of eGFR, among a collection of other biomarkers and demographic information.

Motivated by work in progress by Vironix, we are interested in evaluating questions of physician bias, regional differences, synthetic data generation, and other issues, in the context of this data.

The Iimori dataset provides medical information for 1138 new stage 2–5 CKD patients in the Tokyo metropolitan region. Each patient's age, gender, body mass index, serum hemoglobin, albuminuria, and urine creatinine levels at the time of their first visit were recorded. Other factors provided in the dataset include urine protein-to-creatinine ratio (UPCR), comorbidities (hypertension, diabetes, and cardiovascular disease), and use of RAAS inhibitors, calcium channel blockers, or diuretics. The dataset also provides eGFR measurements every 6 months over 3 years or until the patient is excluded from the study due to death or their eGFR measurement reaching less than 50% of their original eGFR measurement.

Figure 3 illustrates ad-hoc tools we developed to "inject" rare patient trajectories into time series for the downstream goal of better sampling potentially rare events in the synthetic data generation process. One type of rare event is introduction of acute kidney injury (AKI) which, following discussions, is a post-hoc-defined trajectory of eGFR characterized by a sharp drop followed by a return to "normal" behavior. Another type of rare event is a permanent drop after one time point.

With the Iimori dataset, we investigate a type of physician bias based on localized exposure to a global distribution. Therefore, we perform a conceptual experiment with synthetic physicians. Figure 4 illustrates the results. Addressing a hypothesis of physician differences in evaluation of realism of data which is based on regional/professional differences is performed with physicians who are essentially trained on subsets of the same data sampled from different regions in the patient space; see sections 7 and 9 for further discussion.

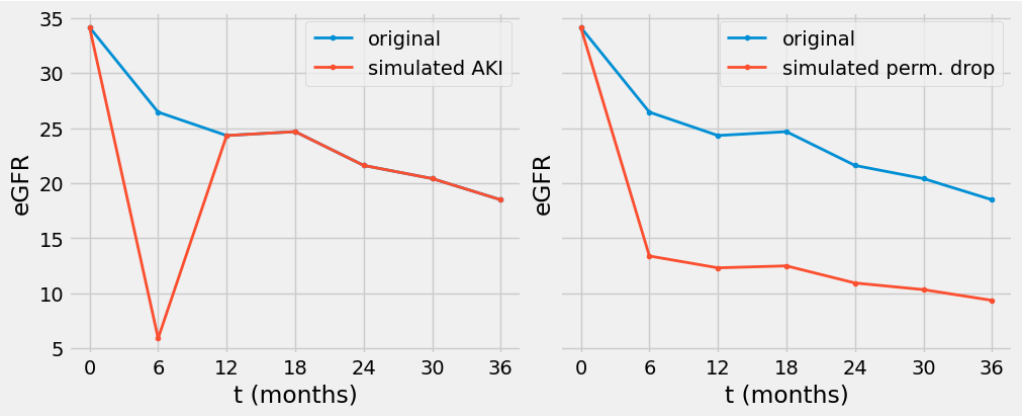


Figure 3. Illustration of an eGFR trajectory over three years from the Iimori dataset ([12]) and two types of forced events. Left: for each timepoint, with probability p, a one-time multiplicative drop is applied to the original eGFR value, aiming to simulate an acute kidney injury (AKI) event. Right: for each timepoint, with probability p, all values past that point are scaled multiplicatively, simulating a permanent acute drop.

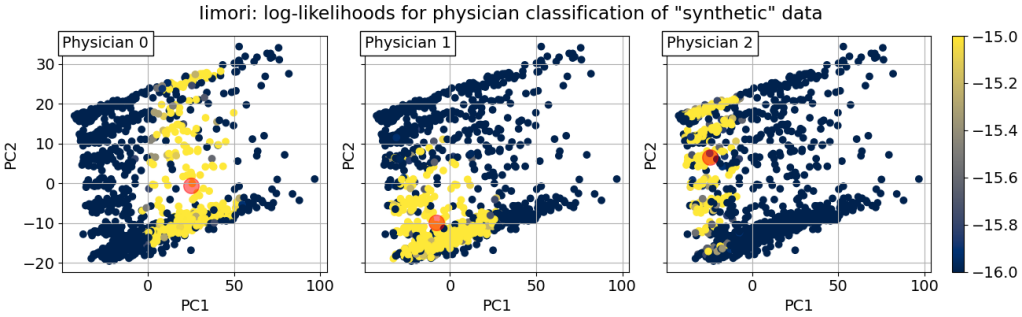


Figure 4. "Physicians" trained on subsets of the data based on Euclidean distance of a 6-dimensional feature space (age, eGFR, eGFR (last visit), slope, observational duration, Cr). Each physician fits a Gaussian distribution to its data, sampled in within a radius of its center (red dot) in the feature space; then is asked to score new data points globally for realism. Data scored below a threshold (blue) are flagged synthetic; those above the threshold (yellow) are accepted as real. Note: PCA projection is for illustration purposes only.

First, for each of the n physicians (indexed by j), physician centers $\mathbf{c}^{(j)}$ in patient parameter space are sampled based on a K-Means clustering, to simplify the task of selecting reasonable choices.[1] Then, a radius r is chosen, and a physician sees one of m patients $x_i \in \mathbb{R}^6$ if $||\mathbf{x}_i - \mathbf{c}^{(j)}||_2 < r$. Then, the collection of data $S_j = \{\mathbf{x}_i : ||\mathbf{x}_i - \mathbf{c}^{(j)}||_2 < r,\ i = 1, \ldots, m\}$ is used to fit a multivariate Gaussian distribution to S_j. With the physician trained, the sample mean and covariance μ_i and Σ_i are used with new

[1] Another low-effort choice would be centering a physician at one of the data points, for example.

Table 1. Description of Data Fields in Vironix's Clinician Survey

Category	**Variable (Description)**
Demographic Data	Age (years) Gender (male/female)
One-time (Initial) Measurements	Body mass index (BMI, kg/m^2) Hemoglobin (Hb, g/dL) Albumin (Alb, g/dL) Creatinine (Cr, mg/dL) Urine protein to creatinine ratio (UPCR, mg/g)
Longitudinal Measurement	eGFR (estimated glomerular filtration rate, measured every 6 months for 3 years, mL/min/1.73m^2)

data x to compute a log-likelihood $s(\mathbf{x};\, \mu_j, \Sigma_j) : \mathbb{R}^6 \to \mathbb{R}$ and is classified synthetic if $s(\mathbf{x};\, \mu_j, \Sigma_j) < t$ for a pre-defined threshold t. In Figure 4, this is implemented with three physicians. Their log-likelihoods for each patient (color, blue to yellow) illustrate that patients near the mode of the physicians' Gaussian models are accepted as real; those far away are called synthetic. We explore extensions of this for more sophisticated machine learning discriminators in section .

6 Clinician Survey

As part of their vetting process, Vironix appealed to a collection of clinicians through a survey. In the interest of establishing a mathematical framework, n clinicians (indexed by j) were presented with the same set of m records (indexed by i). In the particular case under study, $m = 25$ and $n = 5$. The records contain a mixture of real and synthetic data; the record fields are listed in Tab. 1. Clinicians do not know m *a priori*, and the records are presented in randomized order. In this *qualitative realism evaluation*, the clinicians are asked three questions:

(1) Does this patient follow a realistic CKD trajectory? [Yes/No]

(2) On a scale from 1 (least) to 5 (most), how likely do you think it is that this is a real patient?

(3) On a scale from 1 (least common) to 5 (most common), how common would you say this patient trajectory is among those being measured for chronic kidney disease (CKD)?

We initially focus on question #1. Since all the synthetic data statistically "represents a real patient", the clinician should answer "Yes" to question #1 for every case. We think this is a potential source of error and bias, as discussed in §12.

For reasons that will become clear in later sections, it is more mathematically convenient to analyze the number of times the clinicians reject a record. Therefore, we define the *rejection decision* r_{ij} as follows:

$$r_{ij} = \begin{cases} 1, & \text{if clinician } j \text{ answers "No" to question 1 for record } i, \\ 0, & \text{otherwise.} \end{cases} \tag{6.1}$$

Table 2. Possible results of question #1.

	Actual Data Type	
r_{ij}	Synthetic	Real
1	data unconvincing	real data unrecognized
0	data convincing	real data recognized

Given the rejection decisions, we can then construct the *unweighted rejection score* u_i for record i:

$$u_i = \frac{1}{n}\sum_{j=1}^{n} r_{ij} \tag{6.2}$$

Hence $u_i \in [0, 1]$. The goal state for "real" data (real patients or synthetic patients drawn from the distribution) would be $u_i = 0$ while the goal state for "fake" data (synthetic data not drawn from the distribution) would be $u_i = 1$. (In the particular case under study, all records are drawn from the distributions, so the goal is for $u_i = 0$ for all i.)

These scores can be combined into an unweighted rejection vector $\mathbf{u} \in \mathcal{R}^m$:

$$\mathbf{u} = \frac{1}{n} R\mathbf{1}_n, \tag{6.3}$$

where $R \in \mathcal{R}^{m\times n}$ is the matrix of rejection decisions and $\mathbf{1}_n$ is the n-dimensional vector of all ones. Again, given the experimental design for this project, the goal is to drive $\mathbf{u}$ to $\mathbf{0}$.

For the entire data set, there are four possible outcomes, as shown in Tab. 2. Even though it is problematic that a doctor would incorrectly identify real patient data as synthetic (upper right of table), it is superfluous to our purpose here. Since we are concerned only with the identification of synthetic data, we focus most of our efforts on the identification of the synthetic data cases presented.

In particular, we would like to understand why synthetic data which is statistically indistinguishable from real patient data is being flagged by the practitioner. Is there something in the clinician's background which makes them unfamiliar with patients with the synthetic data profile?

To address this question, we first compare the answers to questions #2 and #3, as shown in Fig. 5. At left are the results for all the data, while at right are the results only for the synthetic data. The positive relationship shows that clinicians are more likely to accept the synthetic data as real if they rate the data as "common". In other words, synthetic data is accepted as real if it comports with the clinician's experience.

One more interesting point: note that the bubble for (5,5) is noticeably smaller for the synthetic data than the total data set. In other words, clinicians are very likely to give 5 ratings (high confidence/very common) to real patient data.

There is nuance in these observations. For instance, clinicians may classify trajectories as real or fake dependent purely on whether said trajectories are similar or dissimilar to the clinicians' personal experience. In this case, perceived veracity would be highly dependent on environment factors regarding a clinician's practice, including local geography, culture, and clientele. To some extent, such simplification of the diagnostic process might be seen as a shortcoming of the clinician, a 'bad' bias. On the other hand, it may be that clinicians are better able to distinguish real from fake data when the trajectory is

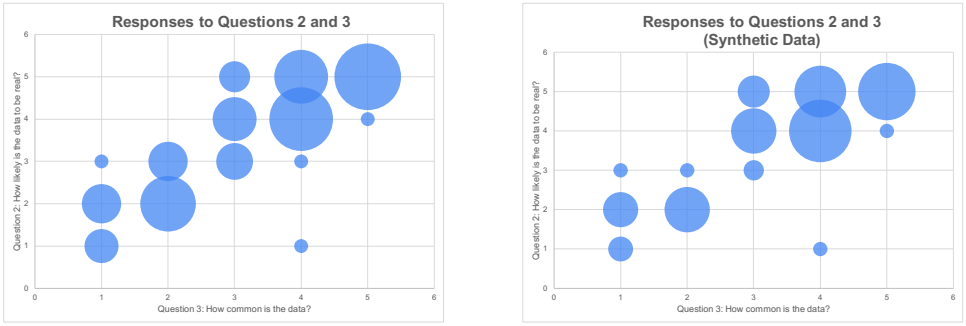


Figure 5. Bubble plot of ratings for question #3 *vs.* question #2. Size of bubble indicates number of responses.

Table 3. Question #3 means for various populations.

	All Patients		**Synthetic Patients**	
r_{ij}	N	Q3 Mean	N	Q3 Mean
0	92	3.80	73	3.78
1	31	1.65	15	1.6

similar to those observed in their practice. In this second case, personal experience influences a clinician's likelihood of identifying data correctly, but does not itself determine the decision. This might be interpreted as a 'good' bias. Both scenarios could lead to an observed correlation between supposed banality of a trajectory and perceived veracity. Proxy models were developed to explore each possibility (Section 9).

Given that the bubble plot seems to indicate a positive relationship between data commonness and acceptability, in Tab. 3 we examine the mean of question #3 for accepted and rejected data. Conclusions of interest:

- The mean of question #3 is much larger for accepted data than rejected data. This result holds for both synthetic patients and all patients.
- The actual means vary very little between the synthetic and all data in each group.
- The rejection rate for synthetic data is 17% (15/73). However, the rejection rate for *real* data is much higher: 46% (16/35). This reinforces the idea that the problem is not whether the data is synthetic or not: rather, it is how it conforms to the clinician's experience.

The tests listed above will help answer the question of *whether* uncommon data is more likely to be rejected by a practitioner. However, it is silent on the question of *why*. Is there something in the records that would cause the record to be rejected?

So as probed in question 3, the question is whether these characteristics are common to a clinician's practice, and if not, will they accept the data anyway?

7 Weighting for Clinician Experience

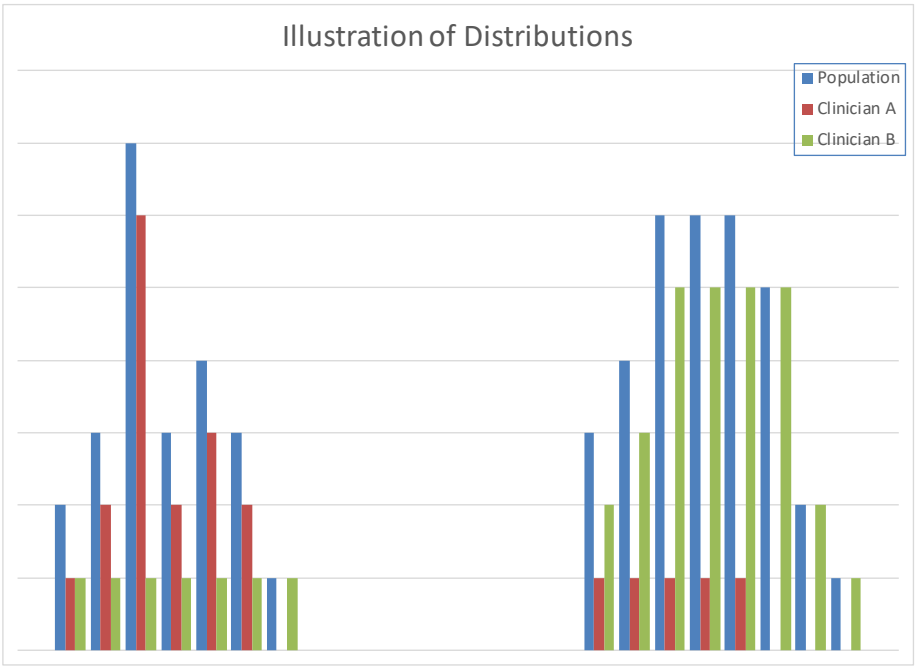


Figure 6. Sketch of effect of clinician experience on variable interpretation. **IMPORTANT:** This is simply an illustration, and does not represent the actual distributions.

To understand how a clinician's experience might affect their acceptance of synthetic data, consider the situation diagrammed in Fig. 6. In blue is a sketch of how a typical biomarker or other variable distribution might look in the general population of late-stage CKD patients. (**IMPORTANT:** This is only an illustration, and should not be considered to be indicative of any true distribution.) In red and green are hypothesized distributions of the same variable for late-stage CKD patients seen by clinicians A and B, respectively.

Given clinician A's experience, they will readily recognize late-stage CKD patients whose parameter values lie in the first peak of the distribution, but would be more unfamiliar with the patients exhibiting the values at right. Hence synthetic data exhibiting such values may be rejected as not representing real patient data. Similarly, clinician B would be more likely to accept patient data in the right portion of the distribution than the left.

7.1 Patient-based Weighting

To formalize this problem, let $P(\mathbf{x})$ be the number of identified CKD patients as a function of some underlying set of biomarker values $\mathbf{x}$. For the purposes of this study, $\mathbf{x}$ may consist of some subset of the fields listed in Tab. 1, perhaps supplemented with some transformation of the data (as described below).

Moreover, let $P_j(\mathbf{x})$ be the number of such patients treated by clinician j. If we make the (hefty) assumption that each patient is treated by exactly one clinician, we have that

$$P(\mathbf{x}) = \sum_{j=1}^{n} P_j(\mathbf{x}), \qquad w_j(\mathbf{x}) = \frac{P_j(\mathbf{x})}{P(\mathbf{x})}. \tag{7.1}$$

Here $w_j(\mathbf{x})$, which we shall call the *weight*, is just the fraction of all patients with characteristic value $\mathbf{x}$ that are seen by clinician j.

However, we must be careful when defining the variable $\mathbf{x}$. For continuous variables (such as the lab results), the number of data points is too small to use continuous variables. Therefore, we would want to bin the quantitative variables into a histogram as in Fig. 6.

Given the weights, we may create a *patient-weighted rejection score* p_i for record i:

$$p_i = \sum_{j=1}^{n} r_{ij} w_j(\mathbf{x}_i), \tag{7.2}$$

where $\mathbf{x}_i$ is the value of $\mathbf{x}$ for record i. Note that (6.2) is a special case of (7.2) where all the weights are $1/n$. Hence $p_i \in [0, 1]$, and again the goal for this study is to have $p_i = 0$.

What is the advantage to this weighting? Consider two clinicians: A and B. A sees many of the patients with profile $\mathbf{x}_i$, and hence has a high weight, while B does not and hence they have a low weight. If B rejects record i, we can discount their opinion because they have little familiarity with that profile. Hence it will not count much towards the rejection score. On the other hand, if A rejects record i, that will carry much more weight since they see many of the patients with that profile.

The score for each record can be generalized into a vector $\mathbf{p}$:

$$\mathbf{p} = \mathtt{diag}(RW), \qquad W \in \mathcal{R}^{n \times m}, \quad w_{ji} = w_j(\mathbf{x}_i). \tag{7.3}$$

7.2 Clinician-based Weighting

The weighting approach described above has objective merit, as it compares the local experiences of each clinician to the characteristic of patients as a whole. However, it still requires knowing characteristics of the entire population of CKD patients. In addition, remember that we are surveying individual clinicians for their conclusions based upon their own local experience. Therefore, we may devise a different weighting system that takes into account how familiar a record is to a clinician's particular experience.

Let T_j be the total number of CKD patients seen by clinician j, and then define

$$f_j(\mathbf{x}) = \frac{P_j(\mathbf{x})}{T_j}. \tag{7.4}$$

Here $f_j(\mathbf{x})$, which we shall call the *patient fraction*, is just the fraction of patients seen by clinician j with characteristic value $\mathbf{x}$.

Given the fractions, we may create a *clinician-weighted rejection score* c_i for record i:

$$c_i = \sum_{j=1}^{n} r_{ij} w_j^*(\mathbf{x}_i), \quad w_j^*(\mathbf{x}_i) = \frac{f_j(\mathbf{x}_i)}{\sum_j f_j(\mathbf{x}_i)}. \tag{7.5}$$

Here the new weights w^* are defined such that they sum to 1, which keeps $c_i \in [0, 1]$, and again the goal for this study is to have $c_i = 0$.

What is the advantage to this weighting? Consider two clinicians: A and B. Most of the patients A sees have profile $\mathbf{x}_i$, and hence A has a high weight. In contrast, few of the patients A sees have profile $\mathbf{x}_i$, and hence B has a low weight. If B rejects record i, we can discount their opinion because they have little familiarity with that profile. Hence it will not count much towards the rejection score. On the other hand, if A rejects record i, that will carry much more weight since they are more likely to see such patients.

The score for each record can be generalized into a vector $\mathbf{c}$:

$$\mathbf{c} = \texttt{diag}(RW^*), \qquad W^* \in \mathcal{R}^{n \times m}, \quad w_{ji}^* = w_j^*(\mathbf{x}_i). \tag{7.6}$$

This framework would allow a description and comparison between weighting schemes, algorithms for discovering/updating weightings based on streaming information $\mathbf{x}_i$, among other applications. We leave further development to possible future work.

8 Data Inference

It may be possible to obtain the medical records corresponding to each clinician participating in the survey, which would allow the computations in §7.2 to be made directly. This couldn't happen on the timescale of the workshop, so instead we work to infer something about their patient populations by comparing the clinical characteristics of accepted and rejected synthetic patients.

We begin by examining the final eGFR value of the patients; the results are shown in Tab. 4. We present the mean of the final eGFR values for patient data that each clinician accepted or rejected. We see a clear pattern: clinicians 1 and 3 accepted data with much lower eGFR values than those they rejected; the remaining three had much smaller deviations between the accepted and rejected data (with clinician 2 rejecting patients with higher final eGFR values). Note also that clinicians 2 and 4 identified every synthetic patient as real.

The results in Tab. 4 are consistent with the fact that progression of CKD is more correlated with *decreases* in eGFR over time, rather than with a particular eGFR value, consistent with what Levey et al. [16] determined in their research in 2003. Hence we now define the following quantity:

$$x = \frac{\Delta(\text{eGFR})}{\Delta t}, \tag{8.1}$$

where the values are taken from the beginning and end of the data collection period. (We use x to indicate that it is a one-dimensional illustration of the $\mathbf{x}$ variable defined in §7.1.)

Table 4. Mean final eGFR value for various populations. n/a represents no patients in this category.

	r_{ij}			
	All Patients		**Synthetic Patients**	
Clinician	Y	N	Y	N
1	15.70	35.75	14.22	37.81
2	27.15	24.43	24.71	n/a
3	26.48	28.00	23.38	28.60
4	26.88	28.20	24.71	n/a
5	24.24	32.64	20.46	39.57

Table 5. x for various populations (value/year). n/a represents no patients in this category.

	r_{ij}			
	All Patients		**Synthetic Patients**	
Clinician	0	1	0	1
1	−4.72	−0.70	−4.96	−0.05
2	−2.83	1.76	−2.78	n/a
3	−3.50	0.14	−3.40	−0.91
4	−2.71	3.43	−2.78	n/a
5	−3.98	0.76	−3.74	0.60

These results are shown in Tab. 5, and they are much clearer. Each clinician was much more likely to accept data which exhibited a large decrease in eGFR (which is indeed correlated with later-state CKD). They rejected data that didn't have large decreases, or even increased. Interestingly, four of the five data records that had increases were real patient data, which is why in the latter column two clinicians didn't reject any of the synthetic patients.

9 Specialty Differences

Another explanation for the discrepancy in clinician evaluations would be their area of training. It has been found that certified nephrologists are more skilled at identifying CKD than primary care doctors [30]. Relatedly, if a clinician sees CKD patients presenting a certain type of profile, how skilled will they be at identifying CKD patients not presenting that profile?

9.1 Proxy A: Commonness of Data

Unfortunately, we do not know the clinicians' background. However, we do know how common they rated each patient's profile. The hypothesis here would be that the more common they rate the patient profile, the more likely they are to correctly distinguish if it is real or synthetic. Given the small number of clinicians surveyed it is hard to draw any firm conclusions, but in Fig. 7 we analyze how *accurate* the clinician is at determining between real and synthetic data. We plot the percentage of records of each type (real and synthetic) classified correctly *vs.* the results of question #3 indicating commonness of

the data. Green bars indicate correctly identified data, while red bars indicate incorrectly identified data. Dark colors are accepted data; light are rejected data.

We begin by examining the real data, at left of each cluster. For low numbers for question #3, the data is uncommon and hence suspect. Hence most real data is flagged as fake (light red). However, for scores of 3 and above, each real data has been accepted (dark green). For the synthetic data, the trend is reversed. For low numbers of question #3, the suspicion about the data leads to correct interpretation of most of the synthetic data (light green). As the numbers increase, the clinicians trust the synthetic data as well, leading to acceptance but misidentification (dark red).

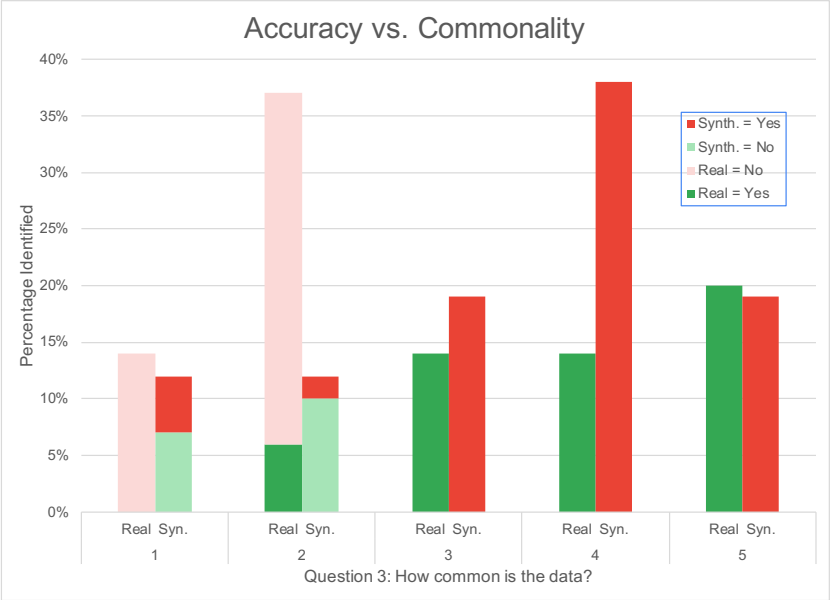


Figure 7. Accuracy of determination *vs.* commonness score. Here green means correctly identified, red means incorrectly identified. Dark means accepted; light means rejected.

9.2 Proxy B: the Boot Approach

Though it may not be possible for the group to obtain the certifications of the clinicians participating in the study, we may model similar situations in the hopes of gaining insight into our current problem. In particular, we use clothing classifications stored in the Fashion MNIST database [32]. Fashion MNIST is a database containing images of various types of clothing, which we treat as analogous to the various types of patients a clinician sees. For this dataset we must also develop synthetic clinicians that will evaluate the data. For this we use what is known as a *discriminator network*. This is a neural

network that is trained by providing it with real and fake data; its objective when training is to correctly discriminate between real and fake data.

We note that this type of synthetic clinician has some important differences from the Gaussian mixture model used in §5, due to what the neural network is optimizing. The more data that the discriminator is trained on, the better it will be at correctly classifying real *vs.* fake data. That means no matter how realistic your synthetic data generation is, if the discriminator is trained equally it will still correctly identify the generated and real data. We think of this type of virtual clinician as like a jeweler. The more diamonds, both real and fake, a jeweler has seen (*i.e.*, their history is the training set) the more likely they are to correctly identify new (test set) diamonds. We believe that the same will also hold for clinicians to some extent - that is, a nephrologist who has seen many patient profiles before will likely be better at correctly discriminating between real and synthetic patients compared to a specialist in another discipline that has seen far fewer CDK patients.

If a discriminator agent is trained with a focus on identifying a particular type of clothing (for instance, a boot), how skilled will it be at identifying other types of clothing, that it has not seen before in its training data?

In particular, we trained three agents on the database:

- One discriminator was trained solely on a dataset of boots.
- Another discriminator was trained solely on a dataset of sandals.
- A third discriminator was trained on both boots and sandals.

We then asked all three agents to identify unseen test images of real and fake boots and sandals in the database.

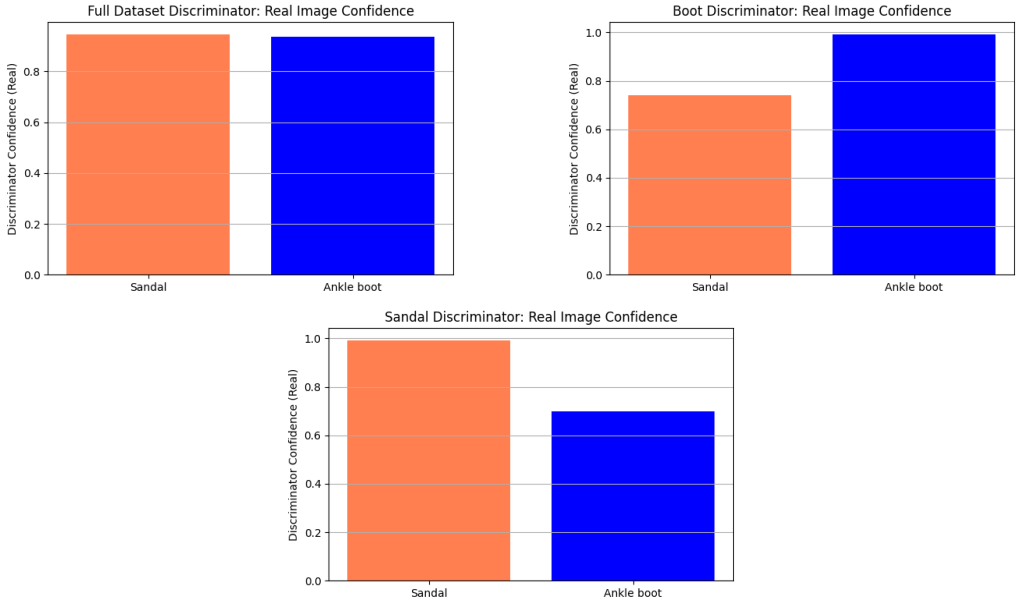


Figure 8. Real image discriminator confidence of the agents.

As shown in Fig. 8, agents trained on one type of footwear identified that type in the testing data with nearly 100% accuracy. They identified the other type of footwear with much less accuracy ($\sim 65-75\%$). The agent trained on both types of footwear identified both roughly equally ($\sim 90\%$), with accuracy between the two limits of the singly-trained agents.

We term the data on which a discriminator is trained its *domain of expertise*. Our goal is to determine how, if we have access only to two individual experts with certain domains of expertise, we can re-weight their answers based on the similarity of a new test sample with the domain of expertise of each discriminator. We think that studying this question also has generalizable benefits to many AI systems. In particular, we may often have a situation in which AI models perform well on subsets of the dataset, and it may be infeasible to train a model on all the data (*i.e.*, due to computational or data privacy constraints).

Figure 9 shows the Principal Component Analysis (PCA) projection of the latent space of the fashion image data. Each point represents an encoded image. We see that the data broadly clusters into boots (blue) and sandals (red): this is a visual representation of the domain-of-expertise differences that may be observed between discriminators (clinicians). When a new test image is provided, we place it in the same latent space (red dot), using for example an autoencoder. This allows us to visualize where the new test point lies within the domain of expertise of our trained discriminators. Our goal is to then reweight the discriminators' scores based on the distance the new data point is from their domain of expertise.

Simply put, we place high weighting on a discriminator's score if the new point is close to ones it has previously seen, and less weighting the further away the test point is from the discriminator's domain of expertise. To classify a new image, we compute its trajectory in the latent space and measure its distance from the mean trajectory of each cluster. These distances are then normalized and used as weights for the agents, allowing us to determine whether the new image is more likely to represent a boot or a sandal.

Translating our results to the medical context, since nephrologists are specially trained to treat CKD, they should be better able to identify real/synthetic CKD patients, as opposed to other types of clinicians. Hence we may be able to introduce an algorithm that weights their rejection decisions more highly than nonspecialists. Unfortunately, given the small number of clinicians in the study, it may not be possible to classify the clinicians by specialty.

9.2.1 *Nick-Nipuni weighting scheme*

A simple weighting scheme that we applied to combine the two discriminator models trained solely on one type of footwear (either boot or sandal), was what we termed the Nick-Nipuni weighting scheme. This is a simple algorithm given by the following steps:

(1) Embed the new test data point into the same latent space as the training set.

(2) Calculate the distance between the test data point and the centroid of the domain of expertise of each discriminator/model. In this case, as a first pass we simply used

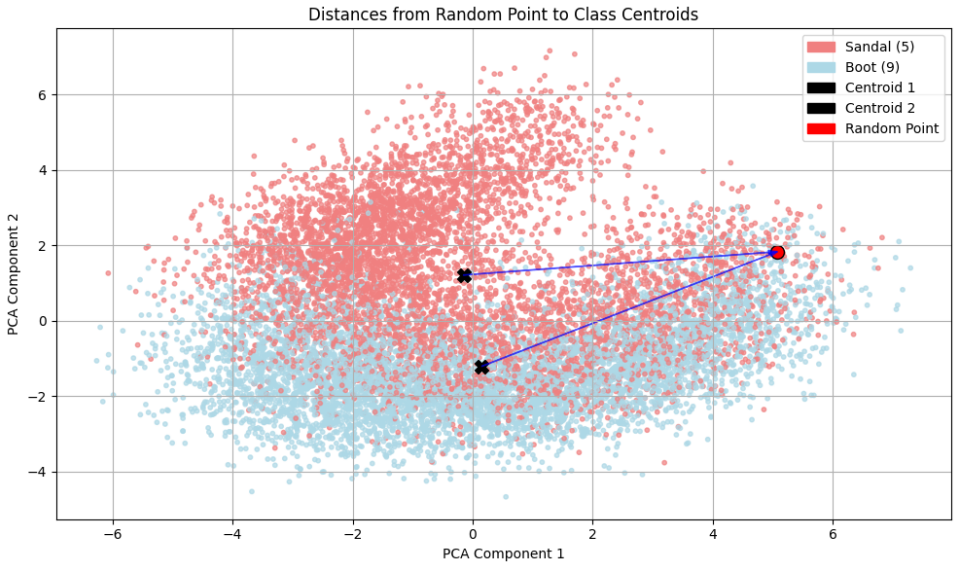


Figure 9. PCA projection of the autoencoder latent space with calculated centroids and arrows showing the Euclidean distance to the random point.

Euclidean distance. These distances form a similarity score, *i.e.*, how similar is a test data point to the domain of expertise for a given discriminator.

(3) Based on the calculated distances, re-weight each discriminator,s score to get a new 'weighted consensus' score.

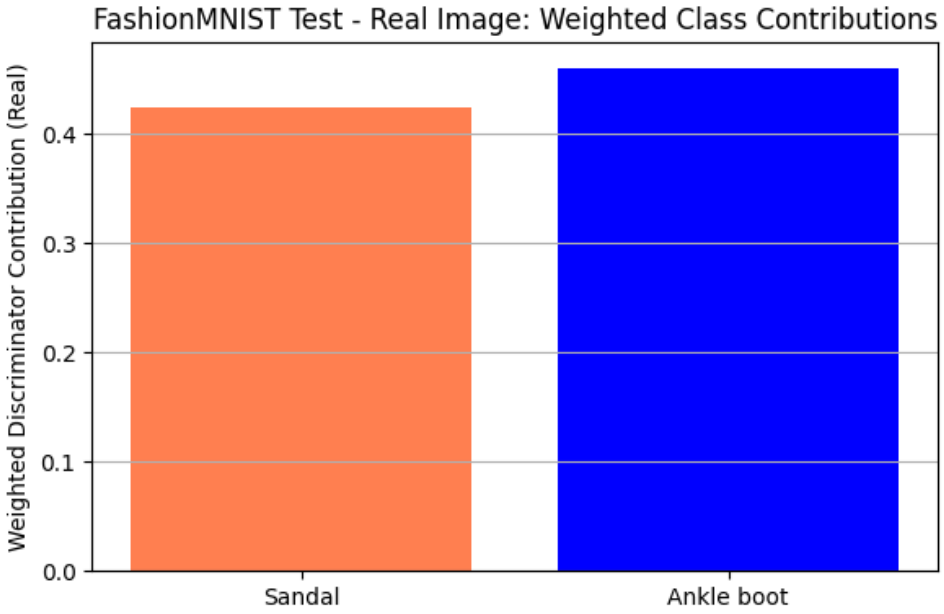


Figure 10. We attempted to combine the scores of two expert discriminators trained solely on either boots or sandals. Unfortunately, the simple Euclidean weighting scheme did not prove successful.

As shown in Fig. 10, this simple approach of Euclidean re-weighting was not successful. We do believe that given more time/thought this method could turn out to be useful in combining discriminator scores. There are a number of areas why it may have failed and potential areas for improvement.

- We used only Euclidean distance to calculate the similarity between new and training data. One could consider many other measures of similarity that are better suited to high dimensional data, *e.g.*, MMD.
- It may be beneficial to introduce a threshold: when a new data point is significantly different from anything the discriminators have been trained on, we may not trust any of their evaluations.
- We calculated the centroid of all the training data and then calculated the distance from that. If the data is not simply clustered, this may not be a good measure to use.

Another ultimate goal of this kind of approach is to introduce some measure of uncertainty in the judgments of experts (be they synthetic or real). The similarity score calculated between a test data point and the training data should somehow inform how much we trust an answer from a given expert. If an expert has seen many data points similar to the new one, we should place high trust in their opinion. However, if an expert has not seen anything like what is shown, weather the data is real or fake, we should be cautious about trusting their evaluation.

9.3 Comparison Between Virtual Clinicians

We have so far considered two potential mathematical formulations of virtual clinicians: Gaussian mixture models (§5) and discriminator networks (§9.2). Given the small number of actual clinicians, these techniques are helpful to test hypotheses about how a cohort of clinicians (or experts in general) may respond when presented with the task of distinguishing real and synthetic data.

We illustrate the key differences between these two types of virtual clinicians by performing a numerical experiment as described below.

9.3.1 *Iimori synthetic data generation*

To compare how different virtual clinicians behave, we generate synthetic patient data from the Iimori dataset using a GAN approach. The results of the data generation are shown in Fig. 11.

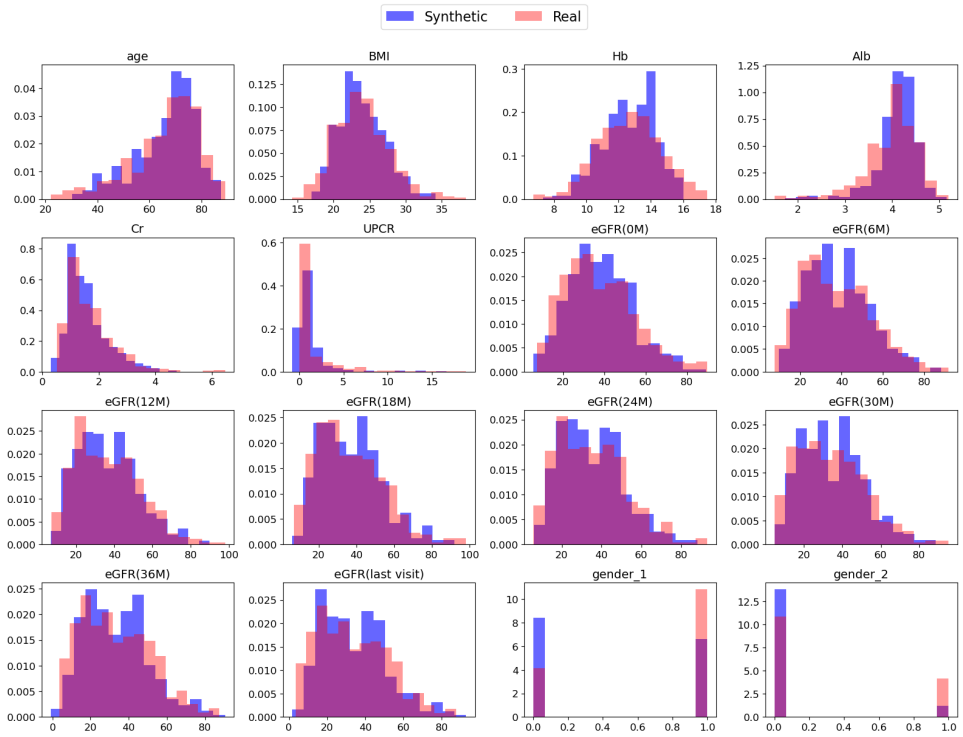


Figure 11. Comparison of distribution for real CDK patient data (red) and synthetic generated patient data (blue). Visually there is good similar between the real and synthetic data.

9.3.2 *Evaluation of virtual clinicians*

Our goal is to evaluate the response of virtual clinicians as the quality of synthetic data generation increases. Hence at each GAN training epoch of training we generate a set

of synthetic patients. We then evaluate our synthetic clinicians by testing them on a set of 100 real and 100 synthetic patient profiles, as shown in Fig. 12. As we train the GAN data generator over more epochs, the data becomes more 'realistic', as shown by a decrease in the MMD between the synthetic and real data (blue line).

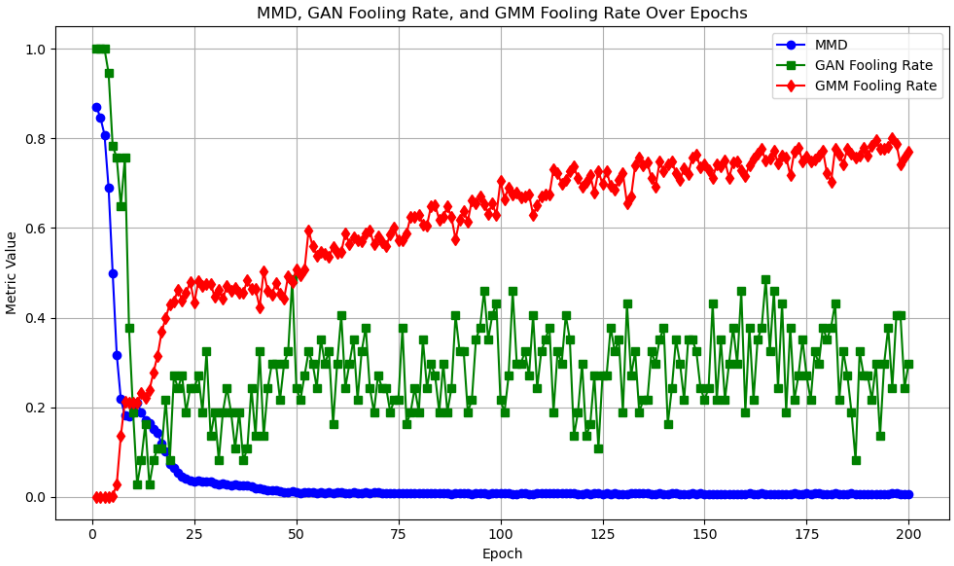


Figure 12. Comparison of the two types of virtual clinician we consider. On the x-axis is number of training epochs on the y-axis is the fraction of synthetic data falsely classified as real. As the number of training epochs increases the synthetic generated data becomes more similar to the real data, this is shown by the MMD decreasing (blue line). As the synthetic and real data becomes more similar the GMM clinician accepts a higher proportion of the data as real. On the other hand the discriminator network is trained in competition with the generator network. Thus as the data gets more realistic the discriminator also gets more capable in distinguishing between real and synthetic data

To generate our virtual clinicans, we apply a Gaussian Mixture Model (GMM). GMMs are one approach to clustering which seeks to describe a collection of data with the assumption that it was sampled from one of a (usually small) number of Gaussian distributions of unknown means and covariances. A key challenge for this space without prior knowledge of the data generation process is inferring the appropriate number of clusters, in addition to the problem of stably identifying the means and covariances themselves. Like many unsupervised methods this has a long history of development of heuristics with goals in mind, often addressing a tradeoff between generalizability and data explainability; fewer centers may generalize but will typically struggle to adequately explain all data. In the opposite extreme, describing the data as n gaussians with covariances $\Sigma = \epsilon I$, with $\epsilon \ll 1$ will fully describe all data but have no utility.

In our application, we apply the GMM framework with only mild by-hand tuning of such parameters. For the clinician application, the hypothetical is as follows. The virtual clinicians accept data as 'real' if it looks similar to what they have seen before. Hence

as the synthetic data generator gets better it has an increased chance to fall within the region the GMM thinks is likely to be real (red line). On the other hand, since the discriminator clinician is trained in competition with the generator network, no matter how similar the synthetic data is to the real data, the virtual clinicians are always able to identify a significant number of synthetic patient profiles.

Unfortunately, due to the small number of real clinician evaluation we have, it is hard to determine which source of more error is dominant. Most likely it is some combination of the two: data that looks more similar to what clinicians have seen before is more likely to be accepted as real, while experts who have seen more CDK patients will also be better at classifying real and synthetic data. Ultimately, these kind of models can help us understand these two different types of bias.

Figure 12 illustrates some results of this process applying a GMM as well as a GAN-based process described above.

10 Interpolation considerations

When examining the data, we found several places where interpolation techniques might be useful:

- We have m records with various values of $\mathbf{x}_i$. How can we then generate realistic profiles for records with values different from that set of m records?
- If the data exhibits clusters clearly identified with a particular state or progression of CKD, how do we characterize a cluster that is somehow "between" them? (See Fig. 13.) Note these clusters can be in symptom space, variable space, behavioral space, geographical location, or some combination of those.

10.1 Proxy C: the Bird Approach

In §9.2, we supposed that clinicians' personal experience prepared them to accurately distinguish real from fake when assessing trajectories common to their practice. This advances the idea of weighing clinician opinions according to expertise. Unfortunately, it is also possible clinicians believe data to be real simply because they have seen similar trajectories before, without critical consideration of the many variables at play. In such a case, clinician opinion is completely determined by their practical experience, which depends on various external factors, including geography (see §3). Hence we require a better understanding of the broad distribution of CKD to compensate for local clinician bias.

This new model aims to understand the interplay between local and global distributions of CKD, as well as how such data is represented in clinicians' 1-5 'commonness' scores (Section 6). Though formulated separately from the boot model, it should be recognized that progressing between global statistics and regional experience is implicitly key to both. In fact, successfully applying conclusions from the multi-clinician approach in the boot model (recall Figure 8) would require knowing each clinician's personal expertise, *i.e.*, their local distribution of CKD trajectories. There is also a question of how local CKD observations impact reported survey scores. After all, Vironix can detect local

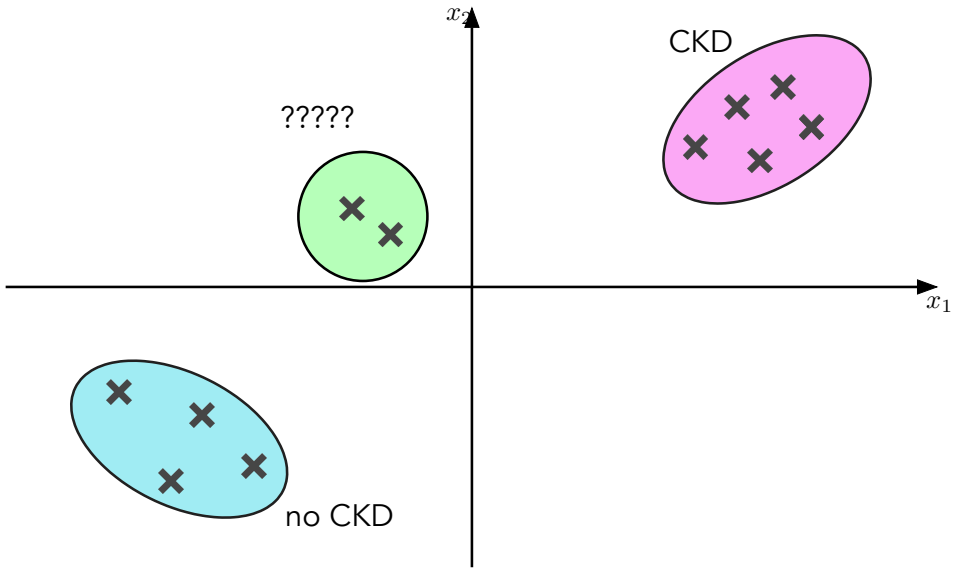


Figure 13. Schematic of data to be interpolated. Here records in the pink and blue regions are given, and records in the green region are to be classified.

distributions only indirectly through these scores. Interpreting survey data requires understanding how to translate between global distributions, local distributions, and survey scores.

To this end, we would like to use clinicians' geographical locations as a proxy for the subpopulation they treat. Unfortunately, privacy protections makes it very difficult to access this sort of data, and hence we are forced to construct a proxy more. From a mathematical perspective, this problems has similar mathematical structure to bird population data, provided by the Cornell Lab of Ornithology through the eBird project [8]. eBird provides birders a platform on which to report local bird sightings. This data is compiled and contextualized by expert ornithologists. In particular, this results in an empirical distribution of bird species across the globe and across avian annual cycles.

Bird data in hand, we propose a model connecting birdwatching to disease observation. Consider the hypothetical placement of a birdwatcher within the continental United States (taken as a subset of the global data available through eBird). This birdwatcher is modeled as observing birds only within a local region, and their perception of bird species abundance is founded on the personal experience of observing in that region.

Figure 14 illustrates an example with 20 birdwatchers, each assigned to a square region measuring 150 units per side. Each unit corresponds to an entry in the frequency matrix and represents approximately 2.5 kilometers horizontally and 3.17 kilometers vertically. In this model, the birdwatcher serves the same mathematical role as a clinician: they report according to their personal, naturally constrained, experience. The various bird species represent different patient trajectories. A birdwatcher's (clinician's) personal impression of a bird's (trajectory's) prevalence determines their opinion of a particular

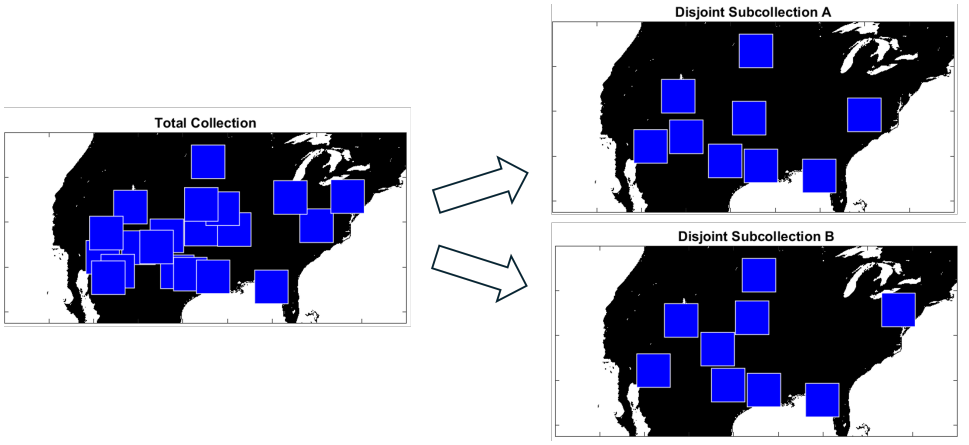


Figure 14. (Left) Random placement of 20 birders with neighborhood radius of 75 units. (Right) Two possible results of pruning that same collection to a disjoint subset.

bird's (trajectory's) likelihood of observation. To believe a hypothetical bird has been observed (based on local experience!) is then equivalent to believing a presented CKD trajectory is real. In this way, the reasoning behind a hypothetical birdwatcher's decision in the artificiality of a claimed observation is analogous to that of a clinician's in deciding the artificiality of patient trajectory. Essentially, the more frequently a birdwatcher sees birds of a certain species relative to others, the more likely they are to expect sightings of that species in the future. Similarly, a clinician who frequently encounters a certain type of CDK may come to expect similar cases and therefore classify a synthetic CDK as real.

To understand our birdwatcher model, and by extension the clinician data, we propose three stages. For the following, the 'global distribution' is a distribution normalized over the entire space of interest: the continental United States in the bird model. A 'local distribution' is a distribution normalized to a subset of the parent space: the region observed by a single birdwatcher in the bird model.

(1) From the known global distribution of birds, construct the local distribution for each hypothetical birdwatcher. From this data, represent individual birdwatchers' perception of relative abundance using a Likert scale. This will illuminate the mechanism by which global statistics determine survey scores.

(2) Given many local distributions, construct a plausible global distribution from which they could have been drawn. This is sometimes called 'the elephant problem' after the old parable. It has long been of general mathematical interest.

(3) Given the abundance ratings (on a Likert scale) of many birdwatchers, construct a plausible global distribution within which those birdwatchers could exist. This stage is the most relevant to Veronix, as it begins with the collected survey data and informs a correction.

These goals are illustrated in the following figure:

It is emphasized that the ideal result of goal (1) is a transformation from the global

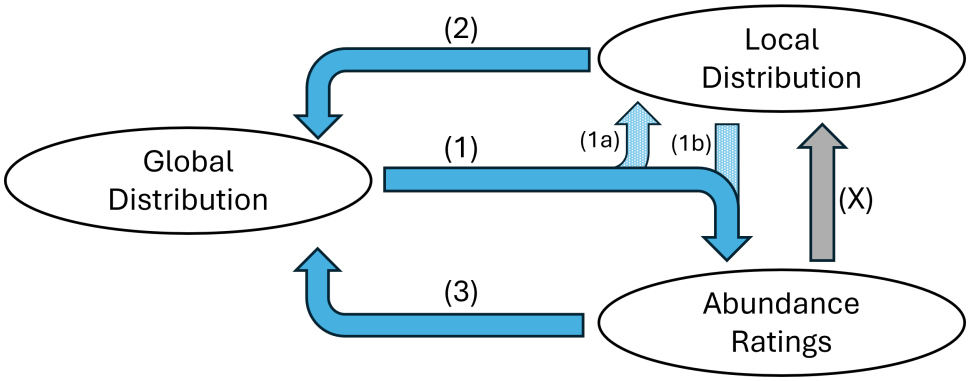


Figure 15. Goals in analysis of bird model.

distribution to local abundance ratings, but that this method involves two steps, (1a) and (1b), so that the local distribution serves as an intermediate calculation.

Note that the ultimate goal (3) is equivalent to resolving (2) and the additional transformation (X) illustrated in the figure. This decomposition may be useful. We expect solutions of (2) to be available in published literature, though time constraints precluded our finding them during the workshop. Conversely, prior experience with probability research suggests (X) to be the most intractable of the proposed goals. We circumvented the challenge by choosing a singularly simple interpolation technique, but expect a more efficient solution will involve significant investigation. For this reason, we focus on framing our goals in terms of (1), (2), and (3).

Goal (1) was approached in a two-part sequence, so-called (1a) and (1b). The goal is to apply knowledge of a global distribution to determine perceived abundance ratings. A notion of local distribution is constructed as an intermediate, hence the proposed decomposition. With respect to the project flow, this pairs (1a) with (2) and (1b) with (X).

The method for constructing local distributions from global distributions in (1a) was inspired by available clinical data. Although clinicians are aware of the number of cases they diagnose, they are presumably ignorant of how those cases are distributed. For example, a clinician might be expected to know that 120 instances of CKD were recently observed in their practice without recognizing that most of these cases came from patients living east of the clinic. This data is presumably available (many medical establishments collect location data from their patients) but is not always processed or available for immediate recall. As such, a clinician's notion of 'local distribution' could reasonably be reduced to a single count for each type of disease trajectory (*e.g.*, high eGFR vs low eGFR—see §4). To incorporate this assumption in the current model, birders are assumed to know only the total count of any given bird in their respective regions. (1a) is therefore resolved by integrating the eBird distributional data over a birder's neighborhood and reporting the counts.

For (1b), we seek to derive abundance ratings from local distributions. To match available survey data, these ratings follow the Likert Scale: scores take values from {1, 2, 3, 4, 5} with greater values corresponding to more frequent observation. Given a set of i

Table 6. Results for Goal 1 using a birder placed in northern Utah (UT) and southeastern New Mexico (NM) with four species of bird under consideration.

Species (i)	**UT** C_i		**UT** S_i	**NM** C_i		**NM** S_i
Western Tanager	0.2512		2	0.0982		3
Mountain Bluebird	0.7444	$\overset{\text{interp.}}{\Longrightarrow}$	5	0.2284	$\overset{\text{interp.}}{\Longrightarrow}$	5
Western Bluebird	0.1429		2	0.0569		2
Eastern Bluebird	0		1	0.0073		1

bird counts for a particular birder, $\{C_i\}_{i=1}^m$, we begin by computing ratios

$$R_i = \frac{C_i}{\sum_{i=m}^n C_k},$$

so R_i is the fraction of total observations by the birder of bird i. Letting max and min denote the indices of the maximal and minimal ratios, respectively, we assign the Likert scores $S_{\max} = 5$ and $S_{\min} = 1$, and determine the cutoffs for different Likert scores by interpolating the ratios and rounding. We chose linear interpolation for simplicity. Table 6 displays the result for a pair of birders, one in northern Utah and the other new southeastern New Mexico.

While this technically resolves (1b), the interpolation step leaves room for further investigation. For instance, linear interpolation results in a distribution on 1-5 different from that observed in the surveys. These surveys exhibited a roughly linear trend across possible scores (Figure 5), but the method proposed here often results in a skew towards lower values of 1 and 2. Logarithmic interpolation may be more appropriate for modeling matching with the survey data.

On a positive note, the proposed method does capture some of the nuance in interpreting distributions from perceived abundance scores. In Table 6, the Western Tanager is rated 2 by the UT birder and 3 by the NM birder. One might suppose that the NM birder must see more Western Tanagers than the UT birder, but this is false. The UT birder observes more birds in total, so the Western Tanager is only *comparatively* less common. In view of the clinician problem, one might suspect that a UT clinician is less likely to identify a 'Western Tanager' trajectory as real because they are relatively rare in their personal practice. Simultaneously, the NM clinician might evaluate that same trajectory as real despite actually observing fewer instances. Herein lies the challenge of using (Likert) scaled local data to test global hypotheses: observers can report only observations relative to their personal environment. This admits cases such as those in Table 6, where different observers offer distinct perspectives and naïve conclusions fail.

To couple with (1b), consider goal (X): reconstruction of a local distribution from abundance ratings. With only the given data, this inverse problem is not well-defined. In fact, solutions are not unique. If a birder observes three birds, providing abundance ratings $\{1, \lambda, 5\}$, the original bird counts (assuming linear interpolation) could be $\{1, \lambda, 5\}$, $\{10, 10\lambda, 50\}$ or even $\{\alpha, \alpha\lambda, 5\alpha\}$ for any positive, real α. To circumvent this obstacle, we assume the total number of birds observed by each birdwatcher is known. This is analogous to supposing clinicians can report their total number of CKD patients, regardless of trajectory. In the case of CKD specialists, this might scale with the size of their practice. While adding assumptions to a model is always frustrating, we hope

Table 7. Results for Goal 3 using results (same experiment) of table 6.

Species (i)	**UT** S_i		**UT** C_i	**NM** S_i		**NM** C_i
Western Tanager	2		0.2277	3		0.1066
Mountain Bluebird	5	$\overset{\text{interp.}^{-1}}{\Longrightarrow}$	0.5692	5	$\overset{\text{interp.}^{-1}}{\Longrightarrow}$	0.1776
Western Bluebird	2		0.2277	2		0.0710
Eastern Bluebird	1		0.1138	1		0.0355

that such data is relatively easy to obtain (at least compared to acquiring the electronic health records of an entire patient pool) so that the method is still relevant to Vironix. In any case, this additional information allows reversion of abundance ratings to local count information. From a set of Likert scores $\{S_i\}_{i=1}^n$, we may calculate counts C_i using the formula

$$C_i = \left(\frac{S_i}{\sum_{j=1}^n S_j} \right) \sum_{k=1}^n C_k.$$

Table 7 shows the results of this procedure when substituting the outputs of Table 6 back into the model.

The choice of interpolation plays an important role here as well. For instance, linear interpolation from 1 to 5 artificially inflates the counts of rare birds while diminishing that of common birds. To see this, note that, in Tables 6 and 7, the model fails to recover the absence of Eastern Bluebirds in UT. Conversely, the prominence of Mountain Bluebirds in UT is cut by nearly a quarter. Substituting the chosen method with linear interpolation from 0 to 4 instead of 1 to 5 would have the opposite effect: artificially lowering and raising the counts of rare and common birds, respectively. We suspect a similar challenge would present itself pairing any interpolative approach with a discrete ranking system. This may suggest the Liker scale to be an inappropriate survey choice if mathematical modeling is the end goal.

All that remains is goal (2). To reiterate, we seek to solve the so-called 'elephant problem' of reconstructing a global distribution from local information. One of the challenges: birder regions may overlap. This is a consequence of our uniformly random placement of birders, an attempt to encode the lack of surveyor control over clinician location. As a result, many methods presented in the literature do not immediately apply. Here we propose our own approach to solving the problem, a procedure that sadly we failed to completely implement in the time allotted to us.

Assume the provision of a collection of birdwatchers, their respective regions, and the number of each kind of bird observed in each region.

(1) Repeatedly prune the collection of regions to find disjoint subcollections.

(2) For a given subcollection, construct a candidate global distribution.

(3) Repeat step (2) with other subcollections, and approximate the actual global distribution with a linear combination of the resulting distributions, ideally weighted to account for the total number of birds represented by each subcollection.

Step (1) resolved the issue of intersection at the cost of data volume. This was successfully implemented, and is illustrated in Figure 14.

Step (2) used the resulting disjoint collection to construct a candidate global distri-

bution. As the method of construction, we chose a variant of two-dimensional linear interpolation. Unfortunately, we struggled to define the problem, especially regarding appropriate boundary conditions from which to base the interpolation. This step was not completed before the end of the workshop. Visually, one could imagine interpolation into the black space of Figure 14.

Step (3) was hoped to offset the loss of information in step (1). Though any single disjoint cover would only cover a portion of the entire region, an aggregate collection of covers should provide information almost everywhere. The notion of compensating for reduced data with repeated sampling was inspired by Monte Carlo methods.

A different method may be more appropriate for step (2). Linear interpolation seemed to falter due to poorly-defined boundary conditions. A Gaussian Mixing Model may be more successful in that regard. The disjoint regions would be approximated as points with a Gaussian fit to each birder. Step (3) would follow as planned. The workshop ended as we attempted to pursue this route. Alternatively, methods from Bayesian statistics may be applicable. The three-step procedure described above may be unnecessary in this case. Our team was not familiar with Bayesian statistics at the outset of this project, so we elected another route, but given the intuitive attraction of a method involving random sampling, a statistical approach may be appropriate.

We attempted to apply Gaussian Mixture Models (GMMs) to reconstruct the global dataset from local observations. However, due to significant differences between the birdwatcher dataset and the datasets commonly used in the literature, as well as fundamental differences in the underlying problem formulation, our approach did not yield satisfactory results. The GMM-based reconstructions diverged considerably from the original data and consistently underestimated bird counts in specific regions. These limitations are likely attributable to both the conceptual mismatch and the limited time available to fully understand and properly implement the GMM framework.

By developing methods to translate from global distributions to abundance scores and abundance scores to local distributions, the bird model succeeded in reproducing phenomena suspected from Vironix's clinician surveys. The first goal illustrated the nuances of using relative abundance to conclude the global distribution, and the third goal illustrated some of the inherent challenges in developing models from survey data. In both cases, an optimal choice of interpolation method may significantly improve the comparison to the clinician data.

Though the project was not completed, we are hopeful the design has potential. A sampling approach to reconstructing global distributions satisfies mathematical instinct, and has precedent in Monte Carlo methods and Bayesian statistics. Upon developing this technique, Vironix could approximate global distributions of CKD trajectories from survey data. This information could be used to correct clinicians' biases when evaluating real and synthetic patient profiles. In conjunction with the boot model, the bird model provides a hopeful start to investigating the relationship between perceived 'commonness' and perceived veracity.

11 Survey Redesign

Our discussions also led to suggestions about how to redesign the survey. First, it is problematic that all the synthetic records presented to the clinicians are drawn from the proper distribution. It is human nature that clinicians, presented with data from a health data company, would expect that they are trying to be "tricked", and would therefore anticipate that some of the data would be fake.

As an analogy, consider the following hypothetical survey. OpenAI presents people with 25 real photographs, asking them to indicate which are real and which are generated with AI. Given that OpenAI is an AI company, the survey takers would understandably be biased to identify at least *one* photograph as AI-generated. Hence, even if a hypothetically unbiased clinician might indeed achieve a rejection score of 0, an actual person might be biased against doing so.

During the week, we vacillated between interpreting question #2 as "real patient" or "real CKD patient", and we suspect that the clinicians may have interpreted the question in various ways as well. Therefore, we propose changing the questions somewhat for increased clarity:

(1) On a scale from 1 (least common) to 5 (most common), how common is this EHR *in your practice*?

(2) On a scale from 1 (least common) to 5 (most common), how common is this EHR *for CKD patients*?

(3) On a scale from 1 (least) to 5 (most), how likely do you think it is that this is a real patient?

(4) On a scale from 1 (least) to 5 (most), how likely do you think it is that this is a real patient *with CKD*?

Note that we draw a distinction between treated patients and CKD patients. Then if a clinician rejects an EHR is fake, we can distinguish between cases where the EHR is perceived to be totally unrealistic (low on all questions) *vs.* those which seem not to be indicative of CKD (low on even-numbered questions). This will also help distinguish between the biases of different types of clinicians since they will see different types of patients.

12 Conclusions and Discussions

Synthetic data offers significant advantages in clinical research, including accelerating study timelines, increasing data accessibility and diversity, and preserving patient privacy. In this study, we assessed the quality of synthetic data through statistical fidelity using the Maximum Mean Discrepancy (MMD) metric and clinical realism, evaluated *via* blinded clinician assessments. A key contribution of our work is the proposal of normalization strategies to account for variability in clinician ratings. This variability may stem from differences in regional population exposure, clinical experience, and institutional practice patterns, all of which can bias realism evaluations. However, a central challenge we encountered was the lack of direct access to real patient data, necessitating

the development of proxy evaluation methods. To address this, we proposed and tested two distinct weighting approaches using surrogate datasets:

(1) The Nick-Nipuni method (applied to the Fashion MNIST dataset), where each clinician is weighted based on their domain specialization and

(2) The Local Exposure method (using the Cornell Lab birdwatching dataset), where weights reflect the relevance of each data point to the clinician's experience.

We then applied both strategies to the Iimori synthetic CKD dataset, which contains trajectory information of CKD patients. Here, we constructed two representative synthetic clinicians one familiar with trajectories showing increasing eGFR values, and another with decreasing eGFR profiles to simulate differential clinical exposure. These findings support the development of more robust, context-aware evaluation frameworks for synthetic clinical data, particularly when real data access is limited. Future work may extend this by incorporating dynamic clinician modeling and broader patient trajectory diversity.

Acknowledgements

We thank Dr. Marina Chugunova and all the faculty and staff at Claremont Graduate University at Claremont for hosting a successful workshop. The workshop was funded in part by the Society of Industrial and Applied Mathematics, as well as industry partners. We also recognize and appreciate the other participants of MPI 2025 for contributing to such a positive and productive research environment. Coding and calculations were assisted by ChatGPT, Gemini, Perplexity, Cursor, StackExchange, and much more.

Author Contributions

The research group shared responsibilities throughout the project, and many team members migrated between tasks. The following assignments should be viewed flexibly: most everyone contributed to every facet of the project to lesser or greater extent. This list only recognizes researchers' primary focus. Moreover, the task of live presentation at the workshop was given, over the course of the week, to Darsh Gandhi (first presentation), Luan Lopes (second presentation), Nicholas Harbour, Adam Petrucci, and Thabo Samakhoana (final presentation).

1 *Introduction*: Hai Van Le, Adam Petrucci

2 *Overview of Synthetic Data Generation*: Hai Van Le, Thabo Samakhoana

3 *Chronic Kidney Disease (CKD)*: Darsh Gandhi

4 *Chronic Kidney Disease from Hong et. al.*: Hai Van Le

5 *Iimori et. al. - Simulated Events, Synthetic Classification*: Manuchehr Aminian, Darsh Gandhi

6 *Clinician Survey*: David Edwards, Hai Van Le

7 *Weighting for Clinician Experience*: David Edwards, Hai Van Le

Nomenclature

Equation numbers where a variable is first defined is listed, if appropriate.

$\mathbf{c}$: vector of clinician-weighted rejection scores (7.5).
f_j: patient fraction (7.4).
i: record index.
j: clinician index.
m: number of records.
n: number of clinicians.
$P(\mathbf{x})$: number of identified CKD patients with characteristic values $\mathbf{x}$.
$\mathbf{p}$: vector of patient-weighted rejection scores (7.2).
R: matrix of rejection decisions r_{ij} (6.1).
T_j: total number of patients seen by clinician j (7.4).
$\mathbf{u}$: vector of unweighted rejection scores (6.2).
W: matrix of patient weights $w_j(\mathbf{x}_i)$ (7.1).
W^*: matrix of clinician weights $w_j^*(\mathbf{x}_i)$ (7.5).
$\mathbf{x}$: vector of characteristic values (6.2).
$\mathbf{1}$: vector of all ones (6.3).

Appendix A Appendix

Table A 1 lists many CKD-related variables clinicians may collect patient data for and the units for each.

Table A 1. Descriptions of CKD-related variables.

Variable	**Description**
age	Age (years)
gender	Gender (male/female/other)
ethnicity	Ethnicity
race	Race
albumin_last	Albumin, last recorded value (g/dL)
albumin_min	Albumin, minimum value (g/dL)
albumin_max	Albumin, maximum value (g/dL)
albumin_median	Albumin, median value (g/dL)
hemoglobin_last	Hemoglobin, last recorded value (g/dL)
hemoglobin_min	Hemoglobin, minimum value (g/dL)
hemoglobin_max	Hemoglobin, maximum value (g/dL)
hemoglobin_median	Hemoglobin, median value (g/dL)
acrenlfail	PMH: Acute and unspecified renal failure (yes/no)
kidnyrnlca	PMH: Cancer of kidney and renal pelvis (yes/no)
nephritis	PMH: Nephritis; nephrosis; renal sclerosis (yes/no)
otdxkidney	PMH: Other diseases of kidney and ureters (yes/no)
bun_last	Blood urea nitrogen (BUN), last recorded value (mg/dL)
bun/creatratio_last	BUN/Creatinine ratio, last recorded value
creatinine_last	Creatinine, last recorded value (mg/dL)
pocbun_last	Point-of-care BUN, last recorded value (mg/dL)
poccreatinine_last	Point-of-care creatinine, last recorded value (mg/dL)
bun_min	BUN, minimum value (mg/dL)
bun/creatratio_min	BUN/Creatinine ratio, minimum value
creatinine_min	Creatinine, minimum value (mg/dL)
pocbun_min	Point-of-care BUN, minimum value (mg/dL)
poccreatinine_min	Point-of-care creatinine, minimum value (mg/dL)
bun_max	BUN, maximum value (mg/dL)
bun/creatratio_max	BUN/Creatinine ratio, maximum value
creatinine_max	Creatinine, maximum value (mg/dL)
pocbun_max	Point-of-care BUN, maximum value (mg/dL)
poccreatinine_max	Point-of-care creatinine, maximum value (mg/dL)
bun_median	BUN, median value (mg/dL)
bun/creatratio_median	BUN/Creatinine ratio, median value
creatinine_median	Creatinine, median value (mg/dL)
pocbun_median	Point-of-care BUN, median value (mg/dL)
poccreatinine_median	Point-of-care creatinine, median value (mg/dL)
egfr_last	eGFR, last recorded value (mL/min/1.73m^2)
egfr(nonafricanamerican)_last	eGFR (non-African American), last value (mL/min/1.73m^2)
egfr(aframer)_last	eGFR (African American), last value (mL/min/1.73m^2)
egfr_min	eGFR, minimum value (mL/min/1.73m^2)
egfr(nonafricanamerican)_min	eGFR (non-African American), minimum value (mL/min/1.73m^2)
egfr(aframer)_min	eGFR (African American), minimum value (mL/min/1.73m^2)
egfr_max	eGFR, maximum value (mL/min/1.73m^2)
egfr(nonafricanamerican)_max	eGFR (non-African American), maximum value (mL/min/1.73m^2)
egfr(aframer)_max	eGFR (African American), maximum value (mL/min/1.73m^2)
egfr_median	eGFR, median value (mL/min/1.73m^2)
egfr(nonafricanamerican)_median	eGFR (non-African American), median value (mL/min/1.73m^2)
egfr(aframer)_median	eGFR (African American), median value (mL/min/1.73m^2)

References

[1] Ahmed Alaa, Boris Van Breugel, Evgeny S. Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 290–306. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/alaa22a.html.

[2] Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela Van Der Schaar. How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pages 290–306. PMLR, 2022.

[3] Peter B Bach, Hoangmai H Pham, Deborah Schrag, Ramsey C Tate, and J Lee Hargraves. Primary care physicians who treat blacks and whites. *New England Journal of Medicine*, 351(6):575–584, 2004.

[4] B. Ballyk. Privacy-preserving generative modelling of longitudinal electronic health records. Master's thesis, University of Oxford, 2024.

[5] Chi D Chu, Neil R Powe, Charles E McCulloch, Deidra C Crews, Yun Han, Jennifer L Bragg-Gresham, Rajiv Saran, Alain Koyama, Nilka R Burrows, Delphine S Tuot, et al. Trends in chronic kidney disease care in the US by race and ethnicity, 2012-2019. *JAMA Network Open*, 4(9):e2127014–e2127014, 2021.

[6] Zachary Dana, Ahmed Ammar Naseer, Botros Toro, and Sumanth Swaminathan. Integrated machine learning and survival analysis modeling for enhanced chronic kidney disease risk stratification, 2024. URL https://arxiv.org/abs/2411.10754.

[7] European Union. General data protection regulation. URL https://gdpr-info.eu/.

[8] D. Fink, T. Auer, A. Johnston, M. Strimas-Mackey, S. Ligocki, O. Robinson, W. Hochachka, L. Jaromczyk, C. Crowley, K. Dunham, A. Stillman, C. Davis, M. Stokowski, P. Sharma, V. Pantoja, D. Burgin, P. Crowe, M. Bell, S. Ray, I. Davies, V. Ruiz-Gutierrez, C. Wood, and A. Rodewald. ebird status and trends. *Cornell Lab of Ornithology*, Data Version: 2023, Release: 2025, Accessed: June 2025, https://doi.org/10.2173/WZTW8903. .

[9] Faris F Gulamali, Ashwin S Sawant, and Girish N Nadkarni. Machine learning for risk stratification in kidney disease. *Current opinion in nephrology and hypertension*, 31(6):548–552, 2022.

[10] Woo Suk Hong, Adrian Daniel Haimovich, and R Andrew Taylor. Predicting hospital admission at emergency department triage using machine learning. *PloS one*, 13(7):e0201016, 2018.

[11] Chi-Yuan Hsu, Feng Lin, Eric Vittinghoff, and Michael G Shlipak. Racial differences in the progression from chronic renal insufficiency to end-stage renal disease in the United States. *Journal of the American Society of Nephrology*, 14(11):2902–2907, 2003.

[12] Soichiro Iimori, Shotaro Naito, Yumi Noda, Hidehiko Sato, Naohiro Nomura, Eisei Sohara, Tomokazu Okado, Sei Sasaki, Shinichi Uchida, and Tatemitsu Rai. Prognosis

of chronic kidney disease with normal-range proteinuria: The CKD-ROUTE study. *PLoS One*, 13(1):e0190493, 2018. ISSN 1932-6203. .

[13] Soichiro Iimori, Shotaro Naito, Yumi Noda, Hidehiko Sato, Naohiro Nomura, Eisei Sohara, Tomokazu Okado, Sei Sasaki, Shinichi Uchida, and Tatemitsu Rai. Prognosis of chronic kidney disease with normal-range proteinuria: The CKD-ROUTE study. *PLOS ONE*, 13(1):1–13, 01 2018. .

[14] Kirsten L Johansen, Glenn M Chertow, Robert N Foley, David T Gilbertson, Charles A Herzog, Areef Ishani, Ajay K Israni, Elaine Ku, Manjula Kurella Tamura, Shuling Li, et al. US renal data system 2020 annual data report: epidemiology of kidney disease in the United States. *American Journal of Kidney Diseases*, 77(4): A7–A8, 2021.

[15] Kamyar Kalantar-Zadeh, Tazeen H Jafar, Dorothea Nitsch, Brendon L Neuen, and Vlado Perkovic. Chronic kidney disease. *The Lancet*, 398(10302):786–802, 2021.

[16] Andrew S Levey, Josef Coresh, Ethan Balk, Annamaria T Kausz, Adeera Levin, Michael W Steffes, Ronald J Hogg, Ronald D Perrone, Joseph Lau, and Garabed Eknoyan. National kidney foundation practice guidelines for chronic kidney disease: evaluation, classification, and stratification. *Annals of internal medicine*, 139(2): 137–147, 2003.

[17] Andrew S Levey, Cassandra Becker, and Lesley A Inker. Glomerular filtration rate and albuminuria for detection and staging of acute and chronic kidney disease in adults: a systematic review. *Journal of the American Medical Association*, 313(8): 837–846, 2015.

[18] MITRE Corporation. Synthetic patient generation. URL https://syntheticahealth.github.io/synthea/.

[19] Hajra Murtaza, Musharif Ahmed, Naurin Farooq Khan, Ghulam Murtaza, Saad Zafar, and Ambreen Bano. Synthetic data generation: State of the art in health care domain. *Computer Science Review*, 48:100546, 2023.

[20] Ahmed Ammar Naseer, Benjamin Walker, Christopher Landon, Andrew Ambrosy, Marat Fudim, Nicholas Wysham, Botros Toro, Sumanth Swaminathan, and Terry Lyons. ScoEHR: generating synthetic electronic health records using continuous-time diffusion models. In *Machine Learning for Healthcare Conference*, pages 489–508. PMLR, 2023.

[21] Afshin Parsa, WH Linda Kao, Dawei Xie, Brad C Astor, Man Li, Chi-yuan Hsu, Harold I Feldman, Rulan S Parekh, John W Kusek, Tom H Greene, et al. Apol1 risk variants, race, and progression of chronic kidney disease. *New England Journal of Medicine*, 369(23):2183–2196, 2013.

[22] Rachel E Patzer and William M McClellan. Influence of race, ethnicity and socioeconomic status on kidney disease. *Nature Reviews Nephrology*, 8(9):533–541, 2012.

[23] C Robertson, A Woods, K Bergstrand, J Findley, C Balser, and MJ Slepian. Diverse patients' attitudes towards artificial intelligence (AI) in diagnosis. *PLOS Digit Health*, 2(5), 2023.

[24] Hector P Rodriguez, Ted Von Glahn, David E Grembowski, William H Rogers, and Dana Gelb Safran. Physician effects on racial and ethnic disparities in patients' experiences of primary care. *Journal of General Internal Medicine*, 23:1666–1672,

2008.

[25] Rudolph A Rodriguez, Saunak Sen, Kala Mehta, Sandra Moody-Ayers, Peter Bacchetti, and Ann M O'Hare. Geography matters: relationships among urban residential segregation, dialysis facilities, and patient outcomes. *Annals of Internal Medicine*, 146(7):493–501, 2007.

[26] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *International conference on algorithmic learning theory*, pages 13–31. Springer, 2007.

[27] Fast Stats. National chronic kidney disease fact sheet, 2017. *US Department of Health and Human Services, Centers for Disease Control and Prevention*, 2017.

[28] Ilya O Tolstikhin, Bharath K Sriperumbudur, and Bernhard Schölkopf. Minimax estimation of maximum mean discrepancy with radial kernels. *Advances in Neural Information Processing Systems*, 29, 2016.

[29] US Department of Health and Human Services. Summary of the HIPAA privacy rule. URL https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html.

[30] Satyanarayana R Vaidya and Narothama R Aeddula. Chronic kidney disease. In *StatPearls [Internet]*. StatPearls Publishing, 2024.

[31] Angela C Webster, Evi V Nagler, Rachael L Morton, and Philip Masson. Chronic kidney disease. *The Lancet*, 389(10075):1238–1252, 2017.

[32] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[33] Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. PhysDiff: Physics-guided human motion diffusion model. In *IEEE International Conference on Computer Vision (ICCV)*, October 2023.

[34] Lianjun Zhang, Chuangmin Liu, and Craig J Davis. A mixture model-based approach to the classification of ecological habitats using forest inventory and analysis data. *Canadian journal of forest research*, 34(5), 2004.