Published in final edited form as:
Phys Lett A. 2010 February 15; 374(9): 1159–1164. doi:10.1016/j.physleta.2009.12.067.

A statistical dynamics approach to the study of human health data: resolving population scale diurnal variation in laboratory data

D. J. Albers and
Department of Biomedical Informatics, Columbia University, 622 W 168th St. VC-5, New York, NY 10032

George Hripcsak
Department of Biomedical Informatics, Columbia University, 622 W 168th St. VC-5, New York, NY 10032

D. J. Albers: david.albers@dbmi.columbia.edu; George Hripcsak: hripcsak@columbia.edu

Abstract

Statistical physics and information theory is applied to the clinical chemistry measurements present in a patient database containing 2.5 million patients' data over a 20-year period. Despite the seemingly naive approach of aggregating all patients over all times (with respect to particular clinical chemistry measurements), both a diurnal signal in the decay of the time-delayed mutual information and the presence of two sub-populations with differing health are detected. This provides a proof in principle that the highly fragmented data in electronic health records has potential for being useful in defining disease and human phenotypes.

Keywords

time-delay dynamics; high-dimensional dynamics; electronic health records; Information theory; statistical mechanics; clinical chemistry; diurnal variation; information theory; mutual information; 05.45.−a; 89.75.−k; 05.45.Tp; 89.70.+c; 89.20.Ff

1. Introduction

Medical research, which aims to improve the prevention, diagnosis, and treatment of disease for individual patients, requires expensive and sometimes risky data collection. Moreover, medical research requires accurate, detailed information about the health state of a large set of patients over a period of time. Said differently, medical research requires data that can resolve human health dynamics on both the individual and population scales. Because health care providers are documenting progressively more of their actions in electronic forms, it *may* be possible to reuse data that are collected during health care and stored in electronic health record (EHR) repositories for medical research. Unfortunately, reusing EHR data brings several challenges: the data are from multiple patients; patients are non-uniformly sampled in time and

over variables (sampling occurs in dense clusters and long time periods pass without measurement); the data are frequently inaccurate or incomplete for logistical and clinical reasons; patients are measured when they are ill, producing sampling bias and correlated sampling times; measured variables are often correlated; the data have very diverse types, ranging from doctor notes (text) to continuous laboratory values; and finally, all patients have unknown stationarity properties. Despite all these issues, usable information is present in the EHR; doctors can read patient records and understand the states and evolution of the patients' health.

From a conceptual standpoint, the EHR is a repository of measurements that contains *both* its own dynamics *and* the dynamics of the subjects it is meant to represent. The inclusion of the dynamics associated with the collection and representation of the data are due to the idiosyncrasies of how the data are collected. Because of this, the problem of reconstructing patients and populations from EHR data has all the ingredients of the fundamental problems in: (i) statistical mechanics (e.g., aggregation of individual patients (microscopic scale) into a population (macroscopic scale) (e.g., disease definitions)); (ii) nonlinear time-series analysis (e.g., most of the data is in the the form of a non-linear time-series, be it lab data or processed text, with very complex bias characteristics); (iii) non-equilibrium dynamics (the stationarity of the patients is unknown, but it is likely that very few of the *measured* (sick) patients are stationary); (iv) nonlinear dynamics in general (these systems can be represented by nonlinear dynamical systems, PDEs, random dynamical systems, stochastic PDEs, etc.).

Whether there exist resolvable signals within EHR data that can be linked to patient or population structure and dynamics is both relatively unknown [3] and contentious [12]. Thus, at its most elemental level, this paper presents a proof in principle: we produce a resolvable signal (based on univariate time-delayed mutual information) from univariate clinical chemistry measurements (CCM) of the patient population in the Columbia University Medical Center (CUMC) - NewYork-Presbyterian EHR using nonlinear physics methodology. More broadly, this paper is directed toward targeting the following problems: (i) can the highly fragmented data present in EHRs be used to resolve a time-dependent, population-wide signal; (ii) are there circumstances when people can be approximated by identical, simple, entities (e.g., like particles); (iii) are there quantifiable laws that govern the dynamics of the humans on a population scale; (iv) can patient sub-populations be inferred from the population-wide, time-dependent signal; (v) if a population-wide signal can be resolved, how representative is this signal of individual patients; and (vi) can the EHR be stratified via a time-based signal. Addressing these questions will provide the data-driven means of defining human *phenotypes* (observable traits), and thus a means of defining disease, health, and human characteristics in general. Achieving the definitions of macroscopic states allows for genetics analysis of complex diseases (e.g., genome wide association studies) where both a complex data-driven genotype and a complex data-driven phenotype are necessary.

Of course this paper does not address all of the above mentioned issues: we are a long way from being able to use the EHR to define phenotypes. Here we focus on four key results. First, time-delayed mutual information (TDMI) can be used to accurately and robustly construct signals representing nonlinear correlation from the complex EHR data source. Second, the *sources* that produce the TDMI signal can be diverse; in particular, the TDMI signal can be composed of intra-source dynamics (e.g., physiology) and inter-source dynamics (e.g., hospital measurement dynamics). Third, homogeneity of a population is difficult to define and has measurable effects on the signal. A single patient can be both heathly and sick, thus allowing each data source to represent multiple statistical states. Moreover, aggregating differently measured differing statistical states can generate statistical correlations in time — hence hospital measurement dynamics can have an effect on the effective TDMI signal. And fourth, qualitative modeling can help provide insight into the statistical structure of the process being

studied. In particular, simple models that reproduce the stylized features of the signals present in real data can be used to explain the source of the signals.

2. Information theoretic analysis of electronic health records

The data source, the CUMC EHR, contains 20 years of data for roughly 2.5 million patients. From this source we target the serum creatinine lab values, of which there are 7.7 million measurements spread over a population of 800, 000 different patients with at least three creatinine values. Most patients have *fewer* than ten creatinine values sampled non-uniformly in time over a matter of years. Nevertheless, the patient with the most values has $\approx 1,300$ spread unevenly over 17 years (this patient's creatinine values are shown in the bottom left plot of Fig. 1). Here we will provide a sketch of the contents of creatinine measurements within the CUMC EHR; however, a complete description of the contents of the EHR, or even relative to the measurements of creatinine, remains an open and difficult problem.

To understand why serum creatinine was chosen, begin by noting that physiologically, creatinine is a waste molecule generated from muscle metabolism; because creatinine is filtered from the blood and then eliminated by the kidneys, serum creatinine levels detail kidney function. Moreover, creatinine is known to have a very weak, but present diurnal signal [4]. Thus, detection of a diurnal signal in creatinine values implies a reasonably fine degree of population resolution. While many creatinine measurements are taken for patients with failing kidneys, often the reason for presence of creatinine measurements are unknown because creatinine is often measured automatically whenever other CCMs (e.g., glucose, sodium, bicarbonate, and phosphorus) are measured. That is, most often clinicians do not order individual laboratory tests but rather order panels that include a variety of tests.

As previously mentioned, the data source is an unfiltered shadow of the population; thus the diversity of the recorded time-series is rather staggering and difficult to overstate. Considering the top left plot of Fig. 1 where the histogram of the mean time between measurements *per patient* is shown, it is clear that while most patients have their measurements taken within a single week, a substantial portion of the population have readings taken considerably less frequently. To crudely characterize non-uniformity between measurements, consider the standard deviation of the time between measurements *per patient*, depicted in the top right plot of Fig. 1. While many patients have narrow standard deviations on the order of days — as can be seen by the sharp peak in the bin representing a one week standard deviation — there is a significant mass of patients whose standard deviation in the time between measurements is on the order of months and years. To further demonstrate this effect, albeit anecdotally, the bottom left plot of Fig. 1 shows three patients with differing numbers of measurements spread over time. The non-uniformity in the time between measurements is largely typical, and relatively pathological. Nevertheless, as can be seen the bottom right plot of Fig. 1, there do exist patients who are measured relatively uniformly in time.

In related work, Komalapriya and colleagues [6] demonstrate both how to, and how well, reconstruction of dynamical systems function when the data consist of many very short time-series. Given the data in Fig. 1, it is likely that a significant portion of the population fits this description. Nevertheless, while restriction to this portion of the population would likely yield less nebulous signals, it is important to note that patients with this sampling signature likely represent a particular subset of patients, not the overall population (i.e., the healthy population will surely be excluded from the onset). Because the goal is to assess the population, including the infrequently measured healthy individuals, we must cope with the all patients regardless of their sampling rates. That said, such work as is contained in Ref. [6] or extensions to it, provide both further evidence that aggregation of many time-series each of which are poorly

sampled can reproduce TDMI signals properly and another tool that may become important as we decipher the EHR.

Due to the non-uniformity in measurement time, the measured creatinine (per patient) can be represented by a time-series $(y(t_1), \ldots, y(t_i), \ldots)$ where y is the creatinine level and t_i is the time it was measured. This yields two time-indexes, *real-time* denoted by t_i, and *sequence-time* denoted by the index i. Thus, there can be two separation times, $\tau = i(t_i) - i(t_{i-\tau})$ and $\delta t = t_i - t_{i-\tau}$; note $\delta t = t_{i+1} - t_i$ is a highly aperiodic positive number. To study the dynamics of creatinine in the population, we use the time-delayed mutual information (TDMI), or the mutual information between all points separated by a *real-time*, δt [5] [11] [2]. Nevertheless, the difference between the TDMI as a function of δt and τ will prove useful when teasing apart the TDMI signal. Define the TDMI by

$$I(Y(t_i), Y(t_{j<i})) = \int p_{y(t_i),y(t_j)} \log \frac{p_{y(t_i),y(t_j)}}{p_{y(t_i)} p_{y(t_j)}} dy(t_i) dy(t_j) \quad (1)$$

The probability density functions (PDF) are estimated with a standard histogram-style measurement; we attained similar results (on fewer patients) when using a kernel density estimator [8] for the PDF. The primary bias of the PDF estimator we used is due to the number of points per bin [10] [1]. To estimate this bias per patient, we randomly permuted the time-ordering of each patient's measurements, thus preserving the distribution of measurements while destroying the time-based information, and then re-estimated the TDMI.

The upper left plot of Fig. 2 represents the mean hourly creatinine variation in the 800, 000 patients whose values are normalized to mean zero and variance one. Note the weak but present diurnal variation in the creatinine values; this shows that despite aggregating over 800, 000 non-uniformly sampled, non-stationary patients over 20 years, an intelligible signal that seemingly agrees with recent results [4] can be reproduced. No small subset of patient records (e.g., the 500 patients with the most readings) can be used to construct this signal, meaning that the aggregation of a large number of patients is necessary to achieve this diurnal signal, given the nature of the sparsely populated individual patient records.

Considering the upper right plot of Fig. 2 carefully, three notable qualitative deductions can be made based on the following observation: the diurnal periodic TDMI signal is *fundamentally* driven by a periodic (in δt) *oscillation in densities* (of y's) that persists within the ensemble of patients integrated over 20 years. *First*, the process generating the TDMI signal cannot be a periodic process plus an arbitrary amount of noise. If the process is a periodic orbit plus any amount of noise, the $Y(t_i)$ versus $Y(t_i - \tilde{t})$ probability density function will contain sharp peaks at the periodic points; and more importantly, the Markov operator moving the densities that generate the TDMI for such a process is a fixed point [7]. *Second*, the peaks in time-delayed correlation are periodic in *separation time* — meaning that raw data separated by 24 hours are more correlated than raw data separated by 2 hours. This is extremely important as it indicates either: (i) the presence of *two distinct intra-patient time-scales*, a slow, daily time-scale that persists over 20 years across patients and is possibly due to the diurnal variation in the population, and a fast, sub-hourly time-scale that goes mostly unresolved, or (ii) the presence of two different populations of patients, each with different TDMI decay rates, one of which is sampled relatively uniformly, and one of which is sampled periodically, and on 24 hour intervals. *Third*, the periodic 24-hour peaks in the TDMI decay smoothly in time — meaning that the correlation between previous time-points is damped away while the correlation window of approximately 16 days persists over the 20 years of data. This indicates that the physiological process is either a chaotic-like signal with phase, or a combination of chaotic/random processes measured non-uniformly.

While we claim that the source of the TDMI signal is an oscillation in densities, determining the source of that oscillation is non-trivial. The evidence in Fig. 2 that supports the hypothesis that one source of diurnal correlation may be inter-patient aggregation can be gleaned by addressing the following: To resolve a reasonably continuous 24-hour signal, there must be some patients measured at all hours of the day; what population gets measured between 10 pm and 6 am? In general, patients that are relatively healthy (i.e., not as ill as others) are sampled at most once a day whereas patients who are more acutely ill are sampled more frequently, including in the middle of the night. Most patients in our sample–which comprises all patients in the CUMC EHR with at least three creatinines–have few creatinine measurements, and most patients are relatively healthy without kidney failure. Healthier patients tend to have fewer measurements.

The effects of this observation can be seen by considering the lower right plot of Fig. 2, which shows the frequency of various δt values between measurements; the peaks at 24 multiples are evident. To use this information to cleave the population in half by hypothesized severity of illness, we removed patients whose measurements correspond to the $\tau < 3$ points; doing this excludes *all* patients with less than three measurements within one to two 24-hour periods. Patients with fewer measurements are likely to be less ill, and are thus removed from the sample. The TDMI curve calculated using this attenuated data set is shown in the upper right plot of Fig. 2. While the diurnal peaks and decay in δt are still present, the diurnal *peaks* have been greatly attenuated by removing the patients we hypothesize are less acutely ill. This observation makes physiologic sense because acutely ill patients are likely to have sudden changes in kidney function, reflected as acute changes in creatinine. Thus, from these calculations, it seems that the diurnal peaks in TDMI are a function of *at least* inter-patient aggregation. Note that the precise number of patients used for the TDMI calculation is 759, 656, resulting in 302, 653, 570 pairs of values. When the $\tau < 3$ patients are filtered out, 302, 827 patients remain, resulting in 283, 037, 944 pairs of values. This implies that there are 19, 615, 626 fewer pairs of values in the $\delta t \leq 2$ graph of the $\tau < 3$ plot. Note that decreasing the number of points per bin will raise the TDMI because of the increase in bias; nevertheless, it is unlikely that sample size has a significant effect on the graph with less patients. Finally, for now we are disregarding more inventive means of combining groups of patients sampled in various ways to yield a periodic decay in the TDMI; given that we are aggregating over 800, 000 patients collected over 20 years, such inventive constructions are seemingly improbable, but certainly not impossible.

While there is physiological evidence and intuition that supports the idea that creatinine may have a diurnal signal in its nonlinear correlation, here we are concerned with what evidence is present in the EHR data for supporting this hypothesis. The first piece of evidence in Fig. 2 that supports the hypothesis that intra-patient physiology can be a source of diurnal correlation is, despite removing all the hypothesized less acutely ill patients and despite the attenuation of the TDMI daily peaks, that the diurnal peaks and decay in δt are still present. Moreover, this decrease in correlation is more consistent with what might be expected physiologically than the sizable peaks present in the entire population. The second piece of evidence that suggests a possible intra-patient source for the diurnal TDMI peaks comes from the lower right plot of Fig. 2, which depicts the TDMI for a single patient (the patient shown in Fig. 1) with 1, 300 points spread over 17 years. Here both the kernel density estimate and the histogram estimates of the TDMI signal are employed; both estimation methods show a rise in the mutual information after the initial trough at six hours and a decay after twenty-four hours. Nevertheless, given only this patient's data, it is unlikely that much more than a single day can be resolved. Moreover, the diurnal signal that is present is relatively faint. These are the reasons why it can be difficult to resolve a signal and why a large population of individuals is necessary for a TDMI estimate.

3. Construction of interpretive models

In the previous section, two different hypothesized sources for the TDMI signal were proposed. In particular, we argue that, to first order, the TDMI signal is a combination of both (i), intra-patient physiology, and (ii) inter-patient aggregation. Nevertheless, it remains to be demonstrated that these sources could, in reality, generate the relevant TDMI signal — a signal that decays and has periodic peaks. To demonstrate that the two proposed sources, physiology and aggregation of differing patients measured differently, can produce the TDMI observed TDMI signal, we will employ two simple models that *qualitatively* represent the claimed TDMI signal sources. In particular, one model will correspond to a qualitative physiology model without aggregation while the other model will correspond to aggregation of two different sources that were measured at different frequencies.

To construct an *intra-patient physiology model* that can reproduce the periodically decaying TDMI signal, define the function that controls the time-evolution of a physiological variable at a particular time, t, of the day by

$$y_{t,\varphi} = f_{\varphi(\tilde{\mathbf{t}})}(\mathbf{a}, \mathbf{x_t}) \tag{2}$$

where: (i) the "slow" time is notated $t \in N$; (ii) the "fast" time is notated by $\tilde{t} = \frac{(t-1)n + j}{n}$ where $j \in [1, n] \subset N$ and n is an integer number of time-grid-points between t and $t + 1$, including the endpoints (t and $t + 1$); (iii) the physiological vector of variables that contribute to the time-evolution of the measurable physiological variable is notated $\mathbf{x_t} \in \mathbf{R^n}$; (iv) the parameters of f are notated $\mathbf{a} \in \mathbf{R^m}$; and (v) the "phase" of f, notated by φ, satisfies $\varphi(\tilde{t}_i) = \varphi(\tilde{t}_j)$ for all i, j such that $|\tilde{t}_i - \tilde{t}_j| \in N$ (φ is identical at every t). The function f can vary functionally with φ — for instance f could vary with daily changes in metabolism. Regardless of whether the functional form or the parameters of f vary over φ, note that an f is attached to every value of $\tilde{t}$, and the $\mathbf{x_t}$-argument of f will vary with φ. To quantify functional variation, define F, which maps the function $f_{\varphi 1}$ to $f_{\varphi 2}$; F is a mapping in function space. This construction implies that for every time grid-point, of which there can be infinitely many, there is a function that guides the CCM physiology at that time grid-point. Moreover, the f at any given time can be coupled to f's at nearby and distant past times via the mapping F. The mapping between the physiological process at time $\tilde{t}_1$ (say, 8 a.m.) and time t_2 (say, 9 a.m.) can have a good deal of physiology incorporated in it. Because we are attempting to make the minimal assumptions necessary to construct a reasonable, interpretable model that captures both the TDMI signal present so that we can determine the essence of the physiological process that yields the TDMI signal, we will refrain from making complex physiological assumptions to construct F. Nevertheless, we will use both physical intuition (or physiological intuition) and the supporting EHR-based evidence to construct the "phase" function, F. An extremely simple F that would satisfy both doctor intuition and the resulting diurnal variation in Fig. 2 is a unimodal function whose mode has a maximum/minimum at the TDMI minimum. Finally, it is worth noting that the intra-patient physiology model formally belongs to one of two classes of systems: time-delay dynamical systems or time-delay differential equations.

Beyond a conception of F being a unimodal function, we must construct an f whose TDMI has peaks that both occur periodically in time and decay to zero. The practical constraints imposed by the periodic, but decaying TDMI signal, include: (i) between every t and $t + 1$ the TDMI is concave up and unimodal; (ii) the peak TDMI values every nt ($n \in N$) time-steps monotonically decreases to zero. Satisfying property (i), or even constructing a TDMI that is periodic is rather simple. However, constructing a TDMI signal that has periodic peaks that *decay* to zero and has a physiological interpretation is more complex. We conceptualize this function as a process

that is strongly affected by the state one period of F ago — for humans, what happened 24 hours ago — and weakly affected by the "out of phase" state (say, 12 hours ago). We will assume F is a linear "sawtooth" between t and $t + 1$:

$$y_{1\le\tilde{t}\le\frac{n+1}{2},\varphi_1(\tilde{t})}=(1-\varphi_1(\tilde{t}))f(a,x_t)+\varphi_1(\tilde{t})f(1,x_t-k) \tag{3}$$

and

$$y_{\frac{n+1}{2}<\tilde{t}<n,\varphi_2(\tilde{t})}=(1-\varphi_2(\tilde{t}))f(a,x_t-k)+\varphi_2(\tilde{t})f^2(x_t-k) \tag{4}$$

(f^2 is f composed with f) where the phase, $\varphi(\tilde{t})$ is defined by:

$$\varphi_1(\tilde{t})=1-\frac{\tilde{t}-1}{\frac{n-1}{2}} \tag{5}$$

and

$$\varphi_2(\tilde{t})=1-\frac{\tilde{t}-1}{n-1} \tag{6}$$

Intuitively, we claim that f at time $\tilde{t} = 1$ (hour 1) is k out of phase with f at time $\tilde{t}=\frac{n+1}{2}$ (hour 12) and that f is back in phase at $\tilde{t} = n$ (hour 24). At intermediate values of $\tilde{t}$, the measurement is defined by a smooth combination of either f at $\tilde{t} = 1$ and $\tilde{t}=\frac{n+1}{2}$ or f at $\tilde{t}=\frac{n+1}{2}$ and $\tilde{t} = n$. These constructions seem reasonable in light of the diurnal variation of mean creatinine value in the population of patients, as can be seen in plot (b) of Fig. 2. For the explicit realization of this construction: set f to be the logistic map, $x_{t+1} = ax_t(1 - x_t)$ with $a = 4.0$; set the number of time grid-points, or n, to five, mimicking the 4 bins per day seen in Fig. 2; and set the phase to be $k = 0.01$. This yields the following explicit set of equations:

$$\begin{aligned}
&y_{1,\varphi_1}=f(4,x_t)\\
&y_{2,\varphi_1}=\tfrac{1}{2}f(4,x_t)+\tfrac{1}{2}f(4,x_t-k)\\
&y_{3,\varphi_1}=f(4,x_t-k))\\
&y_{4,\varphi_2}=\tfrac{1}{2}f(4,x_t-k))+\tfrac{1}{2}f(f(4,x_t))\\
&y_{5,\varphi_2}=f(f(4,x_t))
\end{aligned} \tag{7}$$

thus generating a time-series of:

$$(y_{1,\varphi_1}(t),y_{2,\varphi_1}(t),y_{3,\varphi_1}(t),y_{4,\varphi_2}(t),y_{1,\varphi_1}(t+1),\cdots,y_{4,\varphi_2}(t+M)) \tag{8}$$

In contrast, to specify a *population aggregation model* that will reproduce the periodically decaying TDMI signal, begin with a standard random walk: $z(i) = z(i - 1) + b\zeta_i$, where ζ_i is a Gaussian random variable and b is a real number that controls the TDMI decay rate. Note that the TDMI for z is unbounded in the number of measurements (the TDMI scales with the variance of z); but with finite measurements, the TDMI for z decays smoothly in τ and δ_t and

will thus be qualitatively suitable for the demonstration at hand. Now, to construct a periodic TDMI with a unimodal period, generate two patient populations given by:

$$z(i)=z(i-1)+b_1\zeta_i \tag{9}$$

and

$$\tilde{z}(\tilde{i})=\tilde{z}(\tilde{i}-1)+b_2\zeta_{\tilde{i}} \tag{10}$$

where $\tilde{i}+1-\tilde{i}=10i$ and where $b_2 > b_1$ (thus the TDMI decays slower for the $\tilde{z}$-process). With these two populations, we then treat i as the real-time, and aggregate populations — intuitively this results in one uniformly sampled population and one population (with a lower TDMI decay rate) that is sampled every 10 hours. Hence, this system can represent the construction of a time-series via aggregation of differently measured statistically different sources (e.g., usually, the more acute the illness, the more frequently a patient is measured).

Figure 3 details the periodic decay in the TDMI for both the physiological intra-patient model and the patient aggregation model. Both models reproduce the two key qualitative features: the TDMI smoothly decays to it's relative zero after a finite separation time, and the TDMI decay has an oscillatory nature. Thus, the data-based oscillatory TDMI signal shown in Fig. 2 can be generated by *both* physiology-like models and EHR-patient aggregation models with minimal assumptions.

4. Summary

Upon aggregating 800, 000 patients from a non-uniform population, non-uniformly sampled in time over 20 years, we employed time-delayed mutual information and resolved an intelligible signal that is conceivably representative of patient measurement dynamics within health care, of human physiology, and of various subsets of population according to their predictability. In particular, the two key sources of the observed time-delay correlation signal include: the aggregation of differently measured differing processes and intra-source dynamics. This ability to separate populations and interpret the sources of time-based signals provides hope that in spite of the fact that EHR data is extremely fragmented, issues such as population-wide phenotypes and disease definitions can be extracted. Nevertheless, these results imply that for EHR data to be of use, it must be studied as a natural system in and of itself so as to allow for accurate interpretation of its contents. Aside from the more distant goal of phenotype definition, the analysis here showed: (i) that patient sub-populations of differing health states can be detected and separated; and (ii) population-wide signals from EHR data can be related to simplified qualitative models. For instance, the Lorenz equations, which model convection and are comprised of three coupled ODEs, two detailing the (competing) amplitudes of motion and one detailing the phase between these two amplitudes, have the decaying oscillatory TDMI signal [9] for fundamentally similar (oscillation between densities) qualitative geometric reasons as does the creatinine data and logistic construction. This leads to the conjecture that physiology can be modeled by competing coupled oscillators that do not evolve simultaneously (just like the lobes of the Lorenz equations). Likewise, we conjecture that the presence, dynamics, and measurement of sub-populations can be grossly modeled with aggregated stochastic processes as simple as a random walk.

Acknowledgments

DJA and GH would like to acknowledge the Columbia University Department of Biomedical Informatics data-mining group for helpful discussions; D. Varn for a careful reading of the manuscript; and financial support provided by NLM

grant RO1 LM06910, an award from Microsoft Research for the Phenotypic Pipeline for Genome-wide Association Studies, and a grant from The Smart Family Foundation.

References

1. Albers DJ, Hripcsak G. Estimating correlation and predictive information in electronic health records using mutual information. 2009 submitted.
2. Fraser AM, Swinney HL. Independent coordinates for strange attractors from mutual information. Phys. Rev. A 1986;33:1134–1140. [PubMed: 9896728]
3. Hogan W, Wagner M. Accuracy of data in computer-based patient records. J. Am. Med. Inform. Assoc 1997;5:342. [PubMed: 9292840]
4. Kanabrocki EL, Sothern RB, Sackett-Lundeen L, Ryan MD, Johnson M, Foley S, Dawson S, Ocassio T, McCormick JB, Haus E, Kaplan E, Nemchausky B. Creatinine clearance and blood pressure: a 34-year circadian study. Clin. Ter 2008;159:409–417. [PubMed: 19169600]
5. Kantz, H.; Schreiber, T. Nonlinear Time Series Analysis. 2nd Edition. Cambridge University Press; 2003.
6. Komalapriya C, Thiel M, Ramano MC, Marwan N, Schwarz U, Kurths J. Reconstruction of a system's dynamics from short trajectories. Phys. Rev. E 2008;78:066217.
7. Lasota, A.; Mackey, MC. Chaos, fractals, and noise. Springer-Verlag; 1994.
8. Moon Y-I, Rajagopalan B, Lall U. Estimation of mutual information using kernel density estimators. Phys. Rev. E 1995;52:2318–2321.
9. Pompe B. Measuring statistical dependencies in a time-series. J. Stat. Phys 1993;73:587–610.
10. Roulston MS. Estimating the errors on measured entropy and mutual information. Physica D 1999;125:285–294.
11. Sprott, JC. Chaos and Time-series Analysis. Oxford University Press; 2003.
12. van der Lei J. Use and abuse of computer stored medical records. Meth. Inform. Med 1991;30:79. [PubMed: 1857252]

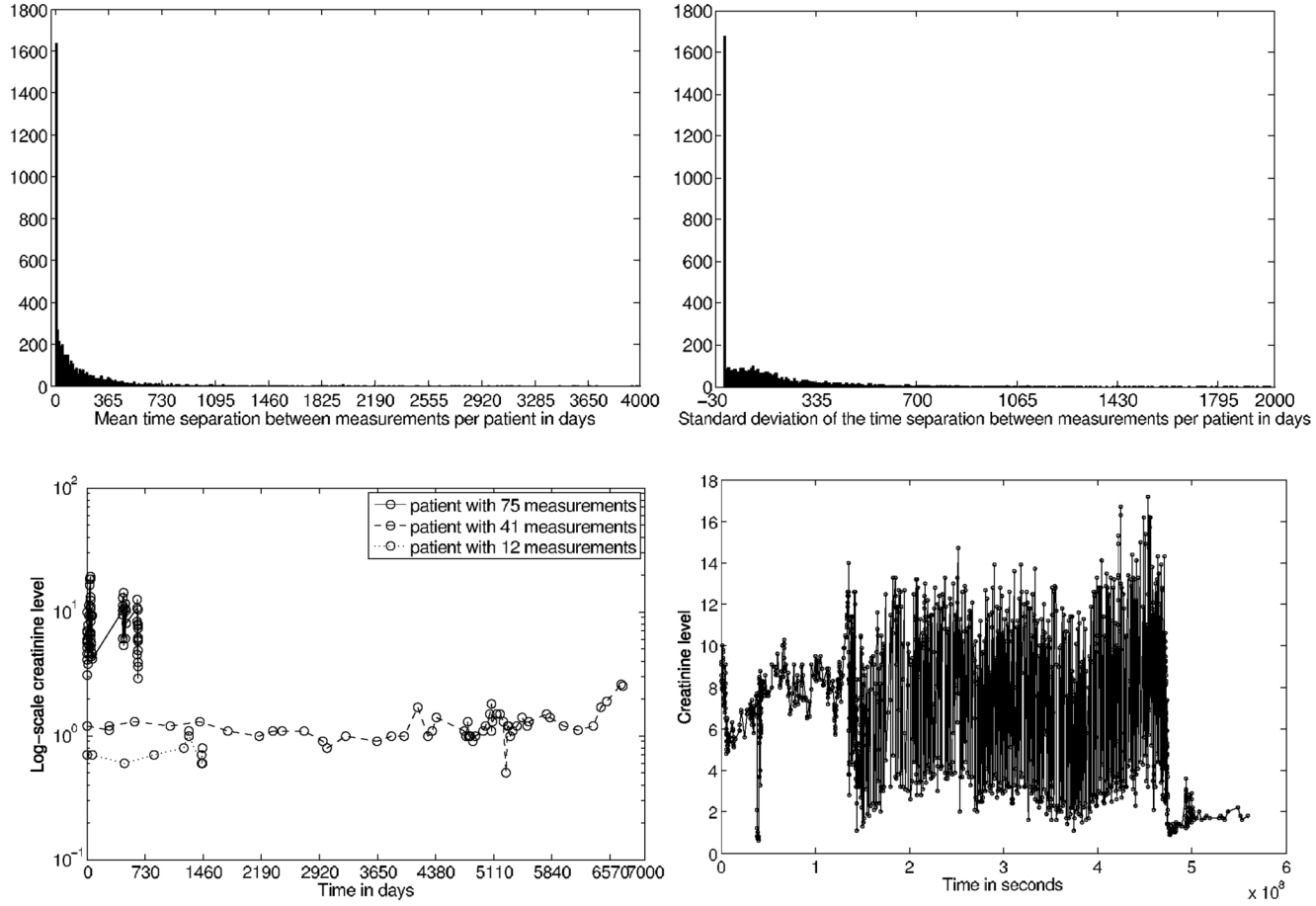

Figure 1.

Upper left (a) depicts a histogram of the mean time between measurements *per patient*; this plot crudely represents the measurement frequency rates. Likewise, upper right (b) shows a histogram of the standard deviation of the time between measurements *per patient*; this plot crudely demonstrates the intra-patient variation in measurement frequency rates. The lower left (c) depicts the creatinine time-series of three different but typical patients. The lower right (d) shows the creatinine time-series for a particularly well sampled patient with 1, 300 creatinine measurements collected over 17 years.

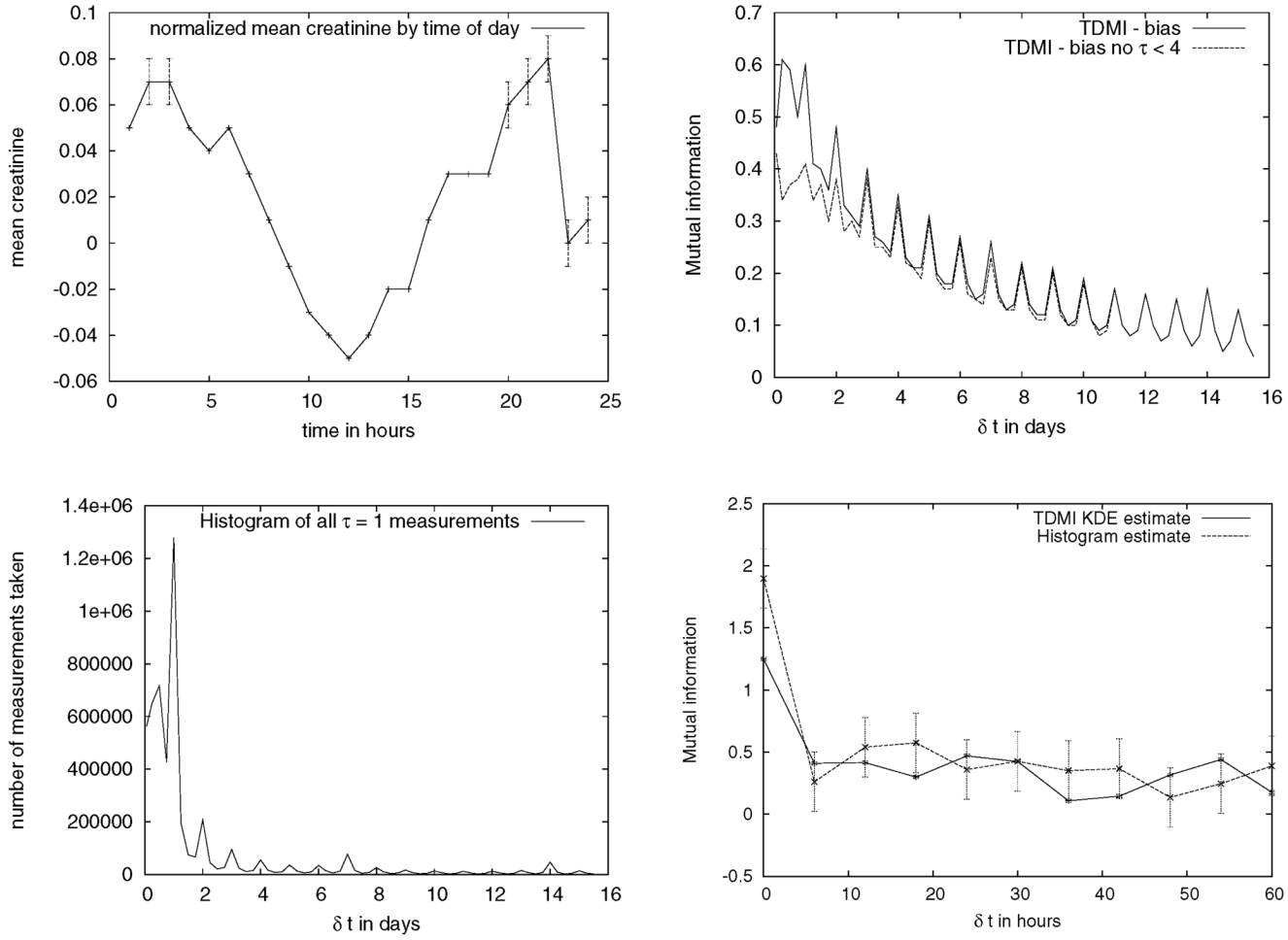

Figure 2.

Upper left (a) depicts the mean *measured* hourly creatinine variation in the 759, 656 (normalized) patients over a day. Note the diurnal variation in the creatinine values. Upper right (b) depicts the variously adjusted time-delay mutual information over a period of 16 days for the 7.7 million creatinine values of 759, 656 patients measured over 20 years. Included are plots of the raw minus the bias TDMI (the corrected) (759, 656 patients, 302, 653, 570 pairs of creatinine measurements), and the TDMI where all $\tau < 3$ $\delta t \leq 2$ points have been excluded (302, 827 patients, 283, 037, 944 pairs of creatinine measurements). Lower left (c) depicts the number of $\tau = 1$ measurements for varying δt. Lower right (d) depicts the TDMI for a *single* patient with 1, 300 creatinine measurements collected over 17 years.

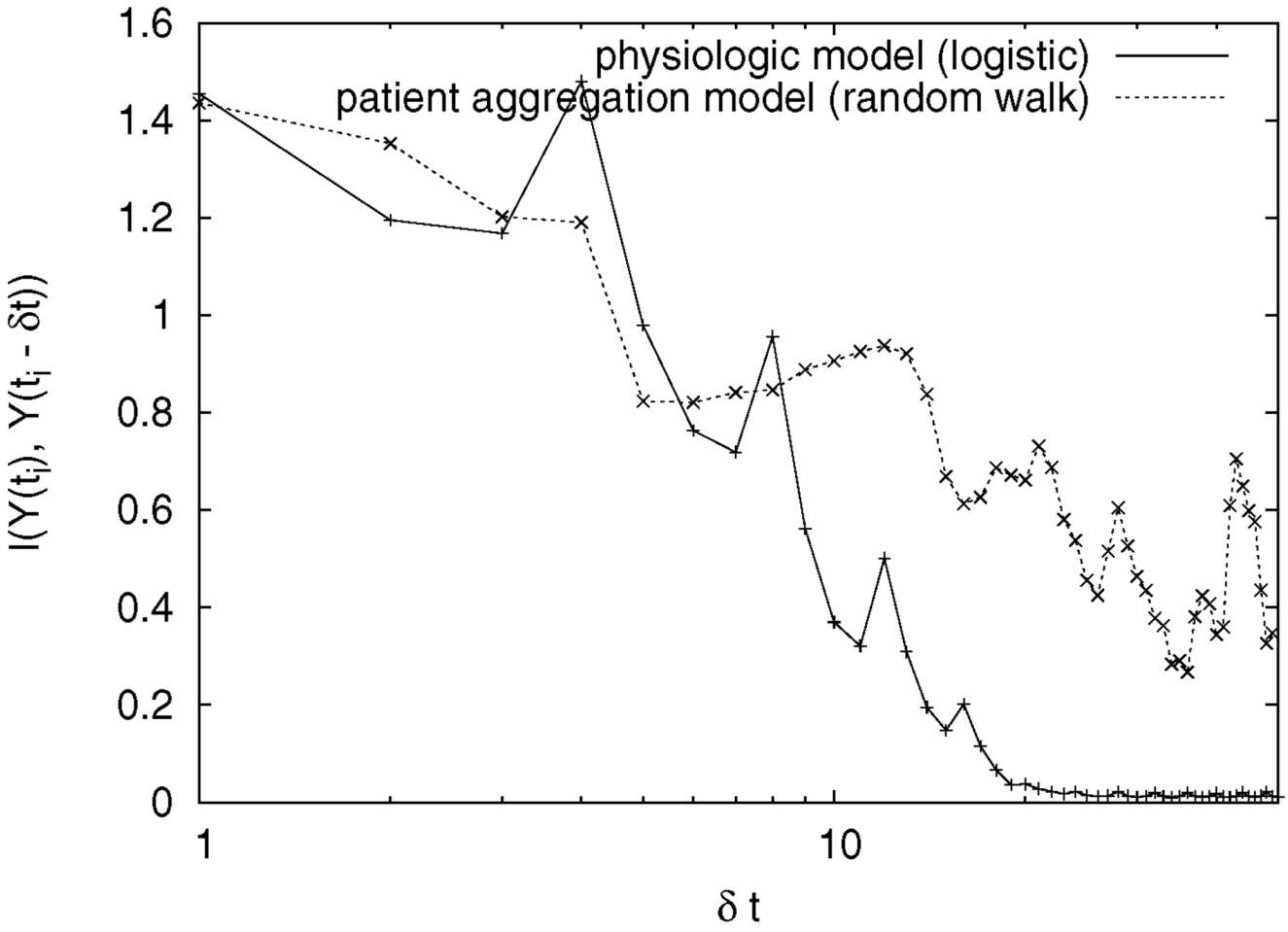

Figure 3.
The time-delay mutual information (as a function of δt) for the intra-patient hypothesis model (interpolated, out-of-phase logistic map) and for the patient aggregation hypothesis model (aggregated random walks with different TDMI decay and sampling rates).