Data Requirements for Model-Based Cancer Prognosis Prediction

Lori A. Dalton[1,2] and Mohammadmahdi R. Yousefi[1]

[1]Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, USA. [2]Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA.

Supplementary Issue: Statistical Systems Theory in Cancer Modeling, Diagnosis, and Therapy

ABSTRACT: Cancer prognosis prediction is typically carried out without integrating scientific knowledge available on genomic pathways, the effect of drugs on cell dynamics, or modeling mutations in the population. Recent work addresses some of these problems by formulating an uncertainty class of Boolean regulatory models for abnormal gene regulation, assigning prognosis scores to each network based on intervention outcomes, and partitioning networks in the uncertainty class into prognosis classes based on these scores. For a new patient, the probability distribution of the prognosis class was evaluated using optimal Bayesian classification, given patient data. It was assumed that (1) disease is the result of several mutations of a known healthy network and that these mutations and their probability distribution in the population are known and (2) only a single snapshot of the patient's gene activity profile is observed. It was shown that, even in ideal settings where cancer in the population and the effect of a drug are fully modeled, a single static measurement is typically not sufficient. Here, we study what measurements are sufficient to predict prognosis. In particular, we relax assumption (1) by addressing how population data may be used to estimate network probabilities, and extend assumption (2) to include static and time-series measurements of both population and patient data. Furthermore, we extend the prediction of prognosis classes to optimal Bayesian regression of prognosis metrics. Even when time-series data is preferable to infer a stochastic dynamical network, we show that static data can be superior for prognosis prediction when constrained to small samples. Furthermore, although population data is helpful, performance is not sensitive to inaccuracies in the estimated network probabilities.

KEYWORDS: cancer prognosis, Bayesian inference, network uncertainty, gene regulatory networks

SUPPLEMENT: Statistical Systems Theory in Cancer Modeling, Diagnosis, and Therapy

CITATION: Dalton and Yousefi. Data Requirements for Model-Based Cancer Prognosis Prediction. *Cancer Informatics* 2015:14(S5) 123–138 doi: 10.4137/CIN.S30801.

TYPE: Original Research

RECEIVED: December 09, 2015. **RESUBMITTED:** February 02, 2016. **ACCEPTED FOR PUBLICATION:** February 07, 2016.

ACADEMIC EDITOR: J. T. Efird, Editor in Chief

PEER REVIEW: Five peer reviewers contributed to the peer review report. Reviewers' reports totaled 1,281 words, excluding any confidential comments to the academic editor.

FUNDING: The work of LAD is supported by the National Science Foundation (CCF-1422631 and CCF-1453563). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: dalton@ece.osu.edu

Introduction

Cancer is characterized by the disruption of molecular pathways that control proliferation, growth, cell death, and energy metabolism and the acquisition of functions that lead to replicative immortality, angiogenesis, metastasis, and evading immune destruction.[1] Cancer prognosis, or the problem of predicting the likely course of disease with or without treatment, is critically important in the treatment of cancer, and particularly difficult given the personalized nature of this disease.

Early tools for cancer prognosis relied primarily on histopathological and morphological features of tissue samples. However, it was soon realized that samples with similar histopathological appearance may correspond to remarkably distinct clinical courses and responses to therapy.[2,3] Beginning in the 21st century, the biomedical community witnessed an explosive success in new technologies that capture the molecular profile of thousands of genes on a single chip, thus facilitating the use of machine learning methods to identify genomic signatures.[4–7]

In 2011, Janet Woodcock, the Director of the FDA Center for Drug Evaluation and Research, estimated that "as much as 75% of published biomarker associations are not replicable."[8] She stated, "This poses a huge challenge for industry in biomarker identification and diagnostics development." Many similar critiques of the current state of affairs have been published in the literature.[9–14] Two potential limitations are small sample size and large measurement noise. Although these problems are beginning to diminish in recent times,[15] there has been little work addressing data requirements or the accuracy of prediction. More fundamentally, progress has been impeded by the complex nature of cancer, namely, the dynamics and uncertainties involved in cell functioning, and the diversity of mechanisms that may drive healthy cells toward a diseased state.[16–18]

The key to progress in precise and personalized diagnosis, prognosis, and treatment strategies for cancer may lie in modeling the topology and dynamics of signaling/regulatory networks and the processes they control, their differences across a population of patients (or cells), responses to intervention, and fully leveraging the vast existing biological prior knowledge and data to build and characterize uncertainty in these models.[19–23] At this time, efforts toward dynamical models for cancer at a gene-regulatory level have only begun to emerge in the computational medicine community.[23,24] A few prognostic and predictive tools have been developed using subnetwork markers that integrate known signaling pathways, gene coexpression networks, protein–protein interaction networks, or somatic mutation profiles.[25–34] However, works in this vein only consider molecular markers with known association, rather than taking full advantage of network topology or regulatory relations. Furthermore, in-depth analysis of these approaches revealed that most do not necessarily outperform traditional tools or provide consistent gene signatures across various studies.[35–37] Recent work[38] provides some hope, concluding that the best performing methods for prognosis prediction in breast cancer incorporate molecular features selected with expert knowledge, as well as both molecular and clinical data.

Our approach to cancer prognosis is illustrated in Figure 1. We assume that an *uncertainty class* of disease models is available, which is a set of models representing plausible variants of aberrant cell functioning. For example, available regulatory pathway knowledge could be used to construct this uncertainty class of models, as proposed in Refs. 39–41. We further assume that gene regulation in each individual in the population is well-modeled by a member of this uncertainty class. There has been some work on gene regulatory models that take into account the stochastic nature of cellular processes, observational noise, and approximation error due to imprecise modeling, model reduction, and latent variables.[42–47] The current work focuses on Boolean network models,[48,49] and to construct the uncertainty class, we optimistically assume that (1) a small set of predictive genes or markers has been identified, (2) there exists a unique healthy network on these genes, which is perfectly known from available prior knowledge, (3) the effect of mutations and abnormalities of this network is perfectly known, and (4) all parameters governing gene interactions are perfectly known. These assumptions are not fundamental to the paradigm in Figure 1 and can be relaxed at the expense of increasing complexity or the size of the uncertainty class. For instance, assumption (2) may be addressed by accounting for incomplete or imperfect knowledge of gene functions and interactions, as in Refs. 39, 41, and 50–54. In general, we only require that the uncertainty class captures the inherent heterogeneity of cancer in the population, including all somatic gene mutations and aberrant pathway functions that may arise in a cancer patient.

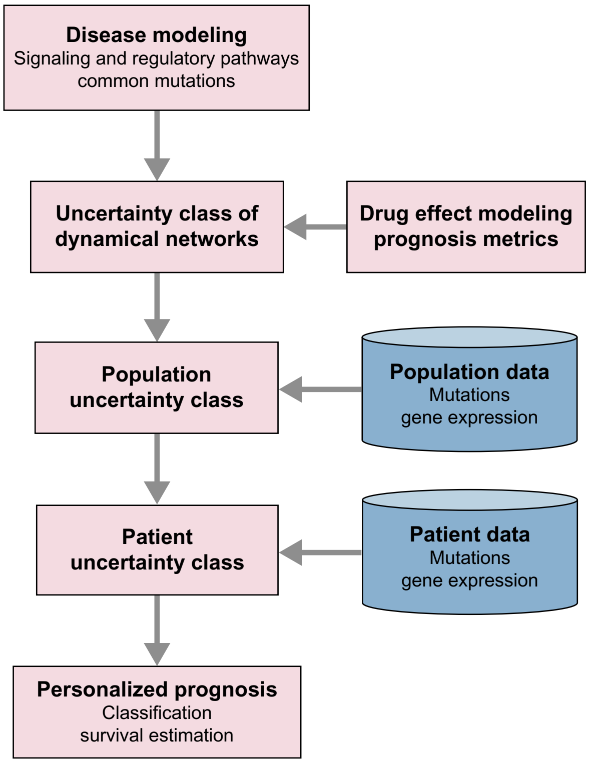

Figure 1. Dynamical modeling for prognosis prediction.

For each network in the uncertainty class, we assign one or more prognostic metrics, which should reflect (genotypic or phenotypic) disease behavior with and without treatment. In our implementation, we assume a treatment strategy has been specified, the resulting network dynamics for each model in the uncertainty class under the influence of the drug is known, and we define prognosis to quantify certain behavior in the gene expression states of a network with and without intervention. This modeling approach was taken in Refs. 55 and 56, where we additionally assumed that a probability distribution of networks in the population is known, eg, that the frequency of common somatic mutations is known. We then formulated optimal Bayesian classification of several prognosis classes based on a single sample of a patient's gene activity profile (GAP). It was shown that accurate prognosis prediction depends greatly on the networks in the uncertainty class and that typically a single sample is not sufficient for accurate prediction.

In this work, we study what experimental measurements are sufficient to predict prognosis reliably. We use population data and a prior to estimate the distribution of networks in the population, patient data to update this to a personalized distribution for the patient, and formulate optimal Bayesian regression of prognosis metrics given this distribution. We address how much data, and what type of data (independent static versus time-series data, and population versus patient data) are needed to predict prognosis. We will show that, under small samples, independent static data is sometimes superior to time-series data for accurate prognosis prediction. Furthermore, although population data is helpful to estimate network probabilities, performance is not sensitive to inaccuracies in these probabilities.

Systems and Methods

Gene regulation model. In this work, we model gene expression (or gene product) levels as binary, where 0 represents inactivation and 1 represents activation of a gene. A *Boolean network* (BN) is a dynamical model for gene regulation, which is characterized by a set of n *nodes*, $v_i \in \{0, 1\}$ for $i = 1, \ldots, n$, representing the expression level of each gene, and a set of n Boolean *predictor functions*, f_i: $\{0, 1\}^n \to \{0, 1\}$ for $i = 1, \ldots, n$, representing gene regulatory relationships. Letting $v_i^k \in \{0,1\}$ be the value of node i at time $k = 0, 1, \ldots$, a GAP is a length-n binary vector, $\mathrm{v}^k = \left[v_1^k, v_2^k, \ldots, v_n^k\right]$, containing the expression level of all n genes at time k. Let $x^k \in S = \{0, 1, \ldots, 2^n - 1\}$ be the integer representation of $\mathbf{v}^k$ given by $x^k = \sum_{i=1}^{n} 2^{n-i} v_i^k$. We call S the *state space* of the network and x^k the *state* of the network at time k. Biologically, a state can be viewed as representing patient phenotype.[21]

Given the current GAP, v^k, the function f_i determines the next value of gene i by $v_i^{k+1} = f_i(\mathrm{v}^k)$. To incorporate known biological regulatory relationships into our gene regulation model, let R be a regulatory matrix with (i, j) entry equal to 1 if gene j activates gene i, –1 if gene j suppresses gene i, and 0 otherwise. We construct predictor functions, f_i, using a "majority voting" rule as follows[52]:

$$f_i\left(\mathrm{v}^k\right) = \begin{cases} 1 & \text{if } \sum_j R_{ij} v_j^k > 0, \\ 0 & \text{if } \sum_j R_{ij} v_j^k < 0, \\ v_i^k & \text{otherwise.} \end{cases}$$

Typically, R is sparse, so f_i depends on a small set of *predictor nodes* rather than the full GAP.

To account for within-model stochasticity, after each time epoch, we assume the current state of each gene in the BN is flipped with probability p. We call this a Boolean network with perturbation (BNp). Any BNp can be represented by a homogeneous Markov chain. In particular, the stochastic process of state transitions is denoted by $\{Z^k \in S: k = 0, 1, \ldots\}$, and the *transition probability matrix* (TPM) is denoted by P, where the (x, y) entry of P, $P(x, y){:} = P(Z^{k+1} = y \mid Z^k = x)$, is the probability of transitioning to successor state $y \in S$ when originating from state $x \in S$ for all time.[49] Gene perturbation in BNps ensures that every state can transition to every other state with positive probability, thus resulting in an ergodic Markov chain with a *steady-state distribution* (SSD), π, representing the long run probability of residing in each state of the network. The probability mass of π at state $x \in S$ is denoted by by $\pi(x)$.

Network uncertainty class. The precise regulatory network in a patient must be initially unknown due to the personalized nature of cancer. We model this with an uncertainty class of possible "cancer" networks that are the products of one or more detrimental, random, and compounding mutations of a nominal "healthy" network, representing normal cell functioning. We also assign prior probabilities to each network in the uncertainty class, Θ, where networks known to occur rarely in the population are assigned lower probability. Thus, we arrive at an uncertainty class, Θ, and probabilities, Λ, where each network $R \in \Theta$ models a cancer subtype and $\Lambda(R)$ models its frequency in the population.

In our implementation, we build Θ by constructing a healthy network from a known healthy regulatory matrix, R^H, and cataloging mutated regulatory networks, each based on one or more modifications of R^H. We require that (1) members of Θ are constructed with a limited number of mutations, and (2) members of Θ must have an undesirable steady-state mass (defined in the "Optimal Bayesian inference with Λ known" section) greater than some threshold. In this way, we ensure Θ has a manageable size and that it contains only mutations with detrimental effects. We also assume that all networks in Θ with i mutations are equally likely and that the sum of their probabilities is proportional to γ^i for some $0 < \gamma \le 1$. Normalizing this distribution to sum to 1, we obtain a valid probability distribution, Λ.

Our prognosis inference methodology does not depend on the method of constructing Θ or Λ. For instance, one might set Λ by drawing from established mutation probabilities in the literature. In the "Optimal Bayesian inference with Λ unknown" section, we discuss the methods of estimating or refining Λ using population data.

Model-constrained Bayesian robust intervention. In practice, the timing and design of intervention are determined by the physician based on experience, knowledge, and data collected from the patient. There may be multiple drugs with many possible effects and side-effects that vary patient to patient. Given an intervention scheme, to infer prognosis we only require that the effects of each drug are known when the network is known, in the sense that the drug-induced dynamics (TPMs or SSDs) are all known. Our analysis is then not particular to the specific intervention strategy used, but for the sake of simulation and illustration, we will employ an optimal robust intervention scheme based on a single idealized drug that affects only a single gene. This is generalized in Ref. 57, which addresses unpredictable effects of a drug on multiple off-target genes, and in Refs. 58 and 59, which address alternate notions of "robust" intervention.

We assume that the expression level of some *control gene* can be altered according to an external input. Let a denote an input taken from a set of actions, $A = \{0, 1\}$, where $a = 0$ models no-intervention and $a = 1$ models toggling the control gene, v_c, corresponding to a node $c \in \{1, 2, \ldots, n\}$. Let $\{(Z^k, A^k) \in S \times A: k = 0, 1, \ldots\}$ denotes the joint stochastic process of both states and actions taken. The TPM for the controlled BNp under action a, P^a, with (x, y) entry $P^a(x, y) = P(Z^{k+1} = y \mid Z^k = x, A^k = a)$, can be easily derived from the TPM of the uncontrolled network, P. In particular, when $a = 1$, each row of the controlled TPM, corresponding to a state x, is found from the row of the uncontrolled TPM corresponding to state x with gene v_c flipped. It is also immediate from the ergodicity of the uncontrolled TPM that the controlled TPM for each $a \in A$ is also ergodic.

The objective of therapy is to alter the behavior of a cancerous cell with a program of treatments that are designed to achieve some desirable effects. To model the effect and performance of therapeutic intervention, let D and U be disjoint subsets of S, which we call *desirable* and *undesirable* states, representing desirable and undesirable phenotypes, respectively. An intervention strategy is a sequence of rules, or a *control policy*, for applying control actions at each time epoch, $k = 0, 1, \ldots, N$, while taking into account previously observed states and actions. The set of all previous states and actions up to time k is denoted by $h^k = (z^0, a^0, z^1, a^1, \ldots, z^k, a^k)$. After observing h^{k-1}, and the current state, z^k, we allow a control policy to implement action $a \in A$ with some prescribed probability $\mu^k(a \mid h^{k-1}, z^k)$. We wish to design an intervention strategy that optimizes the health of the patient by minimizing the long-run average occupation of undesirable states.

This optimization problem has been solved in the framework of optimal Markov decision processes,[60,61] and here we present the salient points. Given a BNp with initial state $Z^0 = x$, and any policy, $\mu = \{\mu^0, \mu^1, \ldots, \mu^N\}$, we may characterize the full joint stochastic process of states and actions for the controlled system, $\{(Z^k, A^k): k = 0, 1\ldots\}$, with a unique probability measure over the space of all trajectories of states and actions, P_x^μ.[60] Minimizing the long-run average occupation of undesirable states is equivalent to minimizing

$$J(x,\mu) = \limsup_{N\to\infty} \mathrm{E}_x^\mu\left[\frac{1}{N+1}\sum_{k=0}^{N}\mathrm{I}(Z^k \in U)\right], \tag{1}$$

where E_x^μ is an expectation relative to the probability measure P_x^μ, and $\mathrm{I}(A)$ is an indicator function equal to 1 if A is true and 0 otherwise.[61] Under initial state $x \in S$, let $J^*(x) = \inf_{\mu\in M} J(x, \mu)$. By definition, a policy μ^* is optimal if $J^*(x) = J(x, \mu^*)$ for all $x \in S$. Not only can one show that an optimal control policy exists, but also it has been shown that an optimal *deterministic* and *stationary* policy exists, which means that at each time epoch, the optimal decision on whether to apply intervention (based on all observed data) can be made deterministically using a fixed rule based on only the current observed state. Moreover, for any stationary policy, μ, $J(x, \mu)$ is invariant to x, and in particular, we may write $J^*(x) = J^*$ for all $x \in S$.[60] The optimal policy can be found using dynamic programming or linear programming.[60,61]

Since the underlying network is not known, we employ Bayesian robust intervention.[51] Using the above theory, we first design an optimal policy for each individual network in the uncertainty class, $R \in \Theta$, and let M be the set of all optimal control policies. We then evaluate the long-run average occupation of undesirable states, $J_R(x, \mu)$, for all combinations of control policies, $\mu \in M$, and networks in the uncertainty class, $R \in \Theta$. Since all policies in M are stationary, as in Eq. (1) we may write $J_R(x, \mu) = J_R(\mu)$ for all $x \in S$. Finally, we define the *model-constrained robust* (MCR) policy, $\mu^\bullet$, to be the policy in M having optimal average performance relative to the prior across the uncertainty class of networks. That is, $\mu^\bullet$ minimizes $\mathrm{E}_\Lambda [J_R(\mu)]$ over all $\mu \in M$, where E_Λ is an expectation relative to the prior, Λ.

Optimal Bayesian inference with Λ known. Assuming the population model to be correct, in this section, we address how one can make optimal inferences regarding prognosis. We will utilize two models for *patient data*, drawn from the individual we wish to prognose: (1) independent observations of the patient's GAP without control, representing sample points of the network state at sufficiently sparse time points to be well modeled as independent and (2) time-series observations of the GAP without control, consisting of sequential observations of the state. In all cases, we denote data of length l by $\mathbf{x} = [x_1, \ldots, x_l]$, with data points $x_i \in S$, $i = 1, \ldots, l$. This data may be based on, for instance, binarized gene expression measurements from microarray or RNA-seq technologies.

The steady-state mass of a network at state x is the long-term probability of visiting state x as the network transitions between states over time. For a given network, $R \in \Theta$, let $\pi_R(x)$ denote the steady-state mass at state x without control. Under independent observations, each GAP is independently drawn from the SSD. Thus, the distribution of x is:

$$f(\mathrm{x} \mid R) = \prod_{i=1}^{l} \pi_R(x_i). \tag{2}$$

Let P_R denotes the TPM without control. Under time-series observations, the probability of transitioning from one GAP to the next is found from the TPM, thus

$$f(\mathrm{x} \mid R) = \pi_R(x_1)\prod_{i=2}^{l} P_R(x_{i-1}, x_i). \tag{3}$$

The prognosis prediction methodology in this study is not dependent on the particular form of the sampling distribution for each network, $f(\mathrm{x}|R)$, but only requires that this data model be specified. One may, for instance, collect observations of the GAP with control, allow for noise in the observed state, allow a continuum of gene expression values, or integrate multiple types of data.

Suppose the uncertainty class, Θ, and network probabilities, Λ, are known, but the specific network at work in a given patient, $R \in \Theta$ is unknown. We wish to estimate the patient's undesirable steady-state mass without control, $\pi_R(U) = \sum_{x\in U} \pi_R(x)$, based on available patient data. This quantity represents the severity of the patient's condition without treatment, where a higher value indicates a worse cancer phenotype. From classical estimation theory, an estimate having minimum mean-square error (MMSE), which we call an "MMSE estimate," is equivalent to the conditional expectation given data. Thus, the MMSE estimate of $\pi_R(U)$ is

$$\hat{\pi}(U) = \mathrm{E}\left[\pi_R(U) \mid \mathrm{x}\right] = \sum_{R\in\Theta} \pi_R(U) f(R \mid \mathrm{x}); \tag{4}$$

where from Bayes rule, and either (2) for independent observations or (3) for time-series observations,

$$f(R \mid \mathrm{x}) = \frac{\Lambda(R) f(\mathrm{x} \mid R)}{\sum_{R' \in\Theta} \Lambda(R') f(\mathrm{x} \mid R')}. \tag{5}$$

The mean-square error (MSE) of $\hat{\pi}(U)$, defined to be MSE $[\hat{\pi}(U)] = \mathrm{E}\left[\left(\hat{\pi}(U) - \pi_R(U)\right)^2 \mid \mathrm{x}\right]$, is

$$\mathrm{MSE}[\hat{\pi}(U)] = \sum_{R\in\Theta} \left(\hat{\pi}(U) - \pi_R(U)\right)^2 f(R \mid \mathrm{x}). \tag{6}$$

The root-mean square (RMS) is the square root of the MSE.

Let $\pi_R^{\bullet}(x)$ denote the steady-state mass at state x under control policy $\mu^{\bullet}$. The undesirable steady-state mass after control, $\pi_R^{\bullet}(U) = \sum_{x\in U} \pi_R^{\bullet}(x)$, represents the severity of a patient's condition with treatment, where a lower value indicates a more effective treatment. The MMSE estimate of $\pi_R^{\bullet}(U)$ given data is

$$\hat{\pi}^{\bullet}(U) = \mathrm{E}\left[\pi_R^{\bullet}(U) \mid \mathrm{x}\right] = \sum_{R\in\Theta} \pi_R^{\bullet}(U) f\left(R \mid \mathrm{x}\right), \tag{7}$$

where $f(R|\mathbf{x})$ is found from Eq. (5). The MSE of $\hat{\pi}^{\bullet}(U)$ is

$$\mathrm{MSE}[\hat{\pi}^{\bullet}(U)] = \sum_{R\in\Theta} \left(\hat{\pi}^{\bullet}(U) - \pi_R^{\bullet}(U)\right)^2 f(R \mid \mathrm{x}). \tag{8}$$

We may also be interested in the shift in undesirable steady-state mass before control to after control. This quantity represents the overall benefit of treatment, where a low value indicates that the drug is less effective. The MMSE estimate of $S_R(U) = \pi_R(U) - \pi_R^{\bullet}(U)$ given data is

$$\hat{S}(U) = \mathrm{E}\left[\mathrm{S}_R(U) \mid \mathrm{x}\right] = \hat{\pi}(U) - \hat{\pi}^{\bullet}(U). \tag{9}$$

The MSE of $\hat{S}(U)$ is

$$\mathrm{MSE}[\hat{S}(U)] = \sum_{R\in\Theta} \left(\hat{S}(U) - S_R(U)\right)^2 f(R \mid \mathrm{x}). \tag{10}$$

In this work, the prognosis of a patient associated with network $R \in \Theta$ is characterized by the values $\pi_R^{\bullet}(U)$ and $S_R(U)$, and the prognoses inferred from data are the values $\hat{\pi}^{\bullet}(U)$ and $\hat{S}(U)$. If we are interested in other metrics for prognosis, for example, toxicity of treatment, side effects, survival time, and quality of life, we only require each network in the uncertainty class, $R \in \Theta$, to be assigned a prognosis score, P_R. Given patient data, the MMSE estimate of P_R becomes $\hat{P} = \mathrm{E}[P_R \mid \mathrm{x}]$, and the MSE of $\hat{P}$ is $\mathrm{MSE}[\hat{P}] = \sum_{R\in\Theta} (\hat{P} - P_R)^2 f(R \mid \mathrm{x})$.

The covariance between two prognostic metrics may also be found. Given data, the covariance between P_R and Q_R is

$$\begin{aligned}\mathrm{Cov}[P_R, Q_R] &= \mathrm{E}\left[\left(P_R - \hat{P}\right)\left(Q_R - \hat{Q}\right) \mid \mathrm{x}\right] \\ &= \sum_{R\in\Theta} \left(P_R - \hat{P}\right)\left(Q_R - \hat{Q}\right) f\left(R \mid \mathrm{x}\right).\end{aligned} \tag{11}$$

For instance, it can be shown that $\mathrm{Cov}\left[\pi_R^{\bullet}(U), S_R(U)\right] = \frac{1}{2}(\mathrm{MSE}[\hat{\pi}(U)] - \mathrm{MSE}[\hat{\pi}^{\bullet}(U)] - \mathrm{MSE}[\hat{S}(U)])$.

Optimal Bayesian inference with Λ unknown. In the previous framework with Λ known, uncertainty stems from patient-to-patient variability in the population, and data from one individual is unhelpful for making inferences on another individual. If Λ is unknown, population data can be used to improve the population model, and therefore, may be indirectly useful to improve inferences about an independent patient.

Suppose the uncertainty class, Θ, is known, but population network probabilities, Λ, are unknown. We assume a Dirichlet prior distribution on Λ, where

$$f(\Lambda \mid \alpha) \propto \prod_{R\in\Theta} (\Lambda(R))^{\alpha(R)-1},$$

and α: $\Theta \to [0, \infty)$ is a set of hyperparameters. This prior places the mean value for $\Lambda(R)$ at

$$\mathrm{E}[\Lambda(R)] = \frac{\alpha(R)}{\sum_{R\in\Theta} \alpha(R)}.$$

The prior is more peaked at the mean when $\sum_{R\in\Theta} \alpha(R)$ is large, and we obtain a uniform prior by assigning $\alpha(R) = 1$ for all $R \in \Theta$. We assume m *population datasets* are available, each dataset being drawn independently from an individual in the population operating under an unknown network drawn from the uncertainty class with unknown probability Λ. The same two models for patient data will be considered for population data. We denote a population dataset with x_j, $j = 1, \ldots, m$, where x_j contains l_j (independent or time-series) points.

Updating the prior with data, from Bayes rule, we obtain a posterior on Λ given by

$$f\left(\Lambda \mid \alpha, \left\{\mathrm{x}_j\right\}_{j=1}^{m}\right) \propto f\left(\Lambda \mid \alpha\right) f\left(\left\{x_j\right\}_{j=1}^{m} \mid \Lambda\right).$$

By enumerating all networks that may govern the population data and applying the law of total probability, we may also write this as

$$f\left(\Lambda \mid \alpha, \left\{x_j\right\}_{j=1}^m\right) \propto f(\Lambda \mid \alpha) \sum_{\substack{R_j \in \Theta \\ j=1,\ldots,m}} f\left(\left\{x_j\right\}_{j=1}^m \mid \left\{R_j\right\}_{j=1}^m\right) f\left(\left\{R_j\right\}_{j=1}^m \mid \Lambda\right),$$

where the sum is over the product space, Θ^m. From our assumption of independence between population datasets, $f\left(\left\{x_j\right\}_{j=1}^m \mid \left\{R_j\right\}_{j=1}^m\right) = \prod_{j=1}^m f\left(x_j \mid R_j\right)$, and from the definition of Λ, $f\left(\left\{R_j\right\}_{j=1}^m \mid \Lambda\right) = \prod_{j=1}^m \Lambda\left(R_j\right)$. Applying our definition of the prior, $f\left(\Lambda \mid \alpha, \left\{x_j\right\}_{j=1}^m\right)$ is also proportional to

$$\sum_{\substack{R_j \in \Theta \\ j=1,\ldots,m}} \prod_{j=1}^m f\left(x_j \mid R_j\right) \prod_{R' \in \Theta} (\Lambda(R'))^{\alpha^*\left(R'; \left\{R_j\right\}_{j=1}^m\right)-1}, \tag{12}$$

where

$$\alpha^*\left(R; \left\{R_j\right\}_{j=1}^m\right) = \alpha(R) + \sum_{j=1}^m \mathrm{I}\left(R_j = R\right). \tag{13}$$

To ease notation, we will sometimes write $\alpha^*\left(R; \left\{R_j\right\}_{j=1}^m\right)$ as simply $\alpha^*(R)$.

The MMSE estimate of $\Lambda(R)$ for any $R \in \Theta$ is the posterior expectation:

$$\begin{aligned}\hat{\Lambda}(R) &= \mathrm{E}\left[\Lambda(R) \mid \alpha, \left\{x_j\right\}_{j=1}^m\right] \\ &= \int \Lambda(R) f\left(\Lambda \mid \alpha, \left\{x_j\right\}_{j=1}^m\right) d\Lambda,\end{aligned}$$

where the region of integration is the space of all valid discrete probability distributions on Θ. Note that since Λ is a valid density, $\hat{\Lambda}$ is also a valid density. From Eq. (12), $\hat{\Lambda}(R)$ is proportional to

$$\int \Lambda(R) \sum_{\substack{R_j \in \Theta \\ j=1,\ldots,m}} \prod_{j=1}^m f\left(x_j \mid R_j\right) \prod_{R' \in \Theta} (\Lambda(R'))^{\alpha^*(R')-1} d\Lambda.$$

Rearranging, we may also write that $\hat{\Lambda}(R)$ is proportional to

$$\sum_{\substack{R_j \in \Theta \\ j=1,\ldots,m}} \prod_{j=1}^m f\left(x_j \mid R_j\right) \int \Lambda(R) \prod_{R' \in \Theta} (\Lambda(R'))^{\alpha^*(R')-1} d\Lambda.$$

It can be shown that

$$\int \prod_{R' \in \Theta} (\Lambda(R'))^{\alpha^*(R')-1} d\Lambda = \frac{\prod_{R' \in \Theta} \Gamma(\alpha^*(R'))}{\Gamma\left(\sum_{R' \in \Theta} \alpha^*(R')\right)},$$

where Γ is the gamma function. Applying this and properties of Γ, $\hat{\Lambda}(R)$ is proportional to

$$\sum_{\substack{R_j \in \Theta \\ j=1,\ldots,m}} \prod_{j=1}^m f\left(x_j \mid R_j\right) \frac{(\alpha^*(R)) \prod_{R' \in \Theta} \Gamma(\alpha^*(R'))}{\left(\sum_{R' \in \Theta} \alpha^*(R')\right) \Gamma\left(\sum_{R' \in \Theta} \alpha^*(R')\right)}.$$

Note that $\sum_{R' \in \Theta} \alpha^*(R') = m + \sum_{R' \in \Theta} \alpha(R')$ is a constant. Hence,

$$\hat{\Lambda}(R) \propto \sum_{\substack{R_j \in \Theta \\ j=1,\ldots,m}} \left(\prod_{j=1}^m f\left(x_j \mid R_j\right)\right) \alpha^*(R) \prod_{R' \in \Theta} \Gamma(\alpha^*(R')).$$

Applying (13) and normalizing, we get $\hat{\Lambda}(R) \propto m(R) + \alpha(R)$, where

$$m(R) = \frac{\sum_{\substack{R_j \in \Theta \\ j=1,\ldots,m}} \prod_{j=1}^m f\left(x_j \mid R_j\right) \left(\sum_{j=1}^m \mathrm{I}\left(R_j = R\right)\right) \prod_{R' \in \Theta} \Gamma(\alpha^*(R'))}{\sum_{\substack{R_j \in \Theta \\ j=1,\ldots,m}} \prod_{j=1}^m f\left(x_j \mid R_j\right) \prod_{R' \in \Theta} \Gamma(\alpha^*(R'))}.$$

Note $m(R) \geq 0$ and $\sum_{R \in \Theta} m(R) = m$ and since $\hat{\Lambda}$ is a valid density,

$$\hat{\Lambda}(R) = \frac{m(R) + \alpha(R)}{m + \sum_{R' \in \Theta} \alpha(R')}. \tag{14}$$

We may interpret $m(R)$ as an effective number of population patients observed from network R. Evaluating $m(R)$ exactly requires evaluating a sum with $|\Theta|^m$ terms, where $|\Theta|$ is the size of the uncertainty class. To approximate $m(R)$, note that

$$m(R) \propto \mathrm{E}_{R_1,\ldots,R_m}\left[\left(\sum_{j=1}^m \mathrm{I}(R_j = R)\right) \prod_{R' \in \Theta} \Gamma(\alpha^*(R'))\right], \tag{15}$$

where in this expectation, each $R_j \in \Theta$ is artificially drawn independently from the density

$$\frac{f\left(x_j \mid R_j\right)}{\sum_{R' \in \Theta} f\left(x_j \mid R'\right)}.$$

This suggests a Monte-Carlo approach to approximate $m(R)$, where we independently draw a large number of m-tuples, $R_1, \ldots, R_m$, from the above density, and we evaluate

$$\left(\sum_{j=1}^m \mathrm{I}(R_j = R)\right) \prod_{R' \in \Theta} \Gamma(\alpha^*(R'))$$

for each m-tuple and each R. We then average across all m-tuples for each R and normalize the resulting values to sum to m, giving approximations of $m(R)$ for all R.

Returning to prognosis prediction, given patient data, x, and population data, $\{\mathrm{x}_j\}_{j=1}^m$, we wish to obtain MMSE estimates of the patient's true prognosis metrics. The MMSE estimate of $\pi_R(U)$ is

$$\begin{aligned}\hat{\pi}(U) &= \mathrm{E}\left[\pi_R(U) \mid \mathrm{x}, \alpha, \{\mathrm{x}_j\}_{j=1}^m\right] \\ &= \sum_{R\in\Theta} \pi_R(U) f\left(R \mid \mathrm{x}, \alpha, \{\mathrm{x}_j\}_{j=1}^m\right),\end{aligned}$$

which is analogous to Eq. (4) with the distribution of networks given patient and population data, $f\left(R \mid \mathrm{x}, \alpha, \{\mathrm{x}_j\}_{j=1}^m\right)$, in place of the distribution of networks given patient data and perfect knowledge of the population network probabilities, $f(R|\mathrm{x})$. From Bayes rule and the law of total probability,

$$\begin{aligned}f\left(R \mid \mathrm{x}, \alpha, \{\mathrm{x}_j\}_{j=1}^m\right) &\propto f\left(R \mid \alpha, \{\mathrm{x}_j\}_{j=1}^m\right) f\left(\mathrm{x} \mid R, \alpha, \{\mathrm{x}_j\}_{j=1}^m\right) \\ &= \int f(R \mid \Lambda) f\left(\Lambda \mid \alpha, \{\mathrm{x}_j\}_{j=1}^m\right) d\Lambda f(\mathrm{x} \mid R) \\ &= \hat{\Lambda}(R) f(\mathrm{x} \mid R).\end{aligned}$$

This is analogous to Eq. (5), where estimates of the network probabilities, $\hat{\Lambda}(R)$ from Eq. (14), replace the unknown true network probabilities, $\Lambda(R)$. To find the MSE,

$$\begin{aligned}\mathrm{MSE}[\hat{\pi}(U)] &= \mathrm{E}\left[(\hat{\pi}(U) - \pi_R(U))^2 \mid \mathrm{x}, \alpha, \{\mathrm{x}_j\}_{j=1}^m\right] \\ &= \sum_{R\in\Theta} \left(\hat{\pi}(U) - \pi_R(U)\right)^2 f\left(R \mid \mathrm{x}, \alpha, \{\mathrm{x}_j\}_{j=1}^m\right).\end{aligned}$$

This is analogous to Eq. (6) with $f\left(R \mid \mathrm{x}, \alpha, \{\mathrm{x}_j\}_{j=1}^m\right)$ in place of $f(R|\mathrm{x})$. Similarly, the MMSE estimate of $\pi_R^\bullet(U)$, the MSE of this estimate, the MMSE estimate of $S_R(U)$, and the MSE of this estimate, are all found using Eqs. (7)–(10), respectively, with $f\left(R \mid \mathrm{x}, \alpha, \{\mathrm{x}_j\}_{j=1}^m\right)$ in place of $f(R|\mathrm{x})$.

Synthetic Data Simulations

In this section, we demonstrate prognosis prediction with synthetically generated networks and data. We begin by describing our method of generating an uncertainty class of networks and associating each network with prognosis metrics. We show how network probabilities may be inferred from population data and show how one may estimate prognosis metrics given patient data.

Network uncertainty class generation. Gene regulatory networks in our synthetic data simulations are based on $n = 6$ genes. To observe a variety of behaviors and results, we generated 100 uncertainty classes of networks, where each uncertainty class represents a different disease, and each network in an uncertainty class represents a subtype of disease. To generate an uncertainty class of networks, we first generate a random *seed* regulatory matrix, R^S, by drawing the number of predictors for each gene uniformly between 1 and 3, thus specifying the number of non-zero entries in each row of R^S, drawing the predictor set for each gene uniformly from the set of all predictor sets of the given size (permitting auto-regulation), thus specifying the location of non-zero entries in each row of R^S, and drawing the regulatory type for each predictor as −1 or 1 with equal probability, thus specifying the values of non-zero entries in each row of R^S. We associate all regulatory matrices with a BNp having perturbation probability $p = 0.01$.

Given R^S, we find a *healthy* network, R^H, by enumerating all regulatory matrices with exactly one edge added or removed in R^S, and selecting the regulatory matrix in this set corresponding to a BNp having minimal undesirable steady-state mass. The uncertainty class, Θ, is then constructed by enumerating all regulatory matrices, R, with one edge added, one edge removed, or one edge added and one edge removed from R^H. We set a *target gene* to v_1, and define undesirable states to be the set of all GAPs such that $v_1 = 0$, corresponding to inactivation of the target gene. Thus, half of the states are undesirable. To ensure that networks in the uncertainty class model detrimental mutations of R^H, we remove networks with undesirable steady-state mass less than that of R^H, or less than the average undesirable mass of all networks with a single mutation, from the uncertainty class.

To set the true probability of each network in the uncertainty class, $\Lambda(R)$ for $R \in \Theta$, we set γ to 0.5. Thus, the sum of probabilities of networks with 1 mutation (one edge added or removed) is $0.5/0.75 = 2/3$, and the sum of probabilities of networks with 2 mutations (one edge added and one edge removed) is $0.5^2/0.75 = 1/3$. All networks with the same number of mutations are equally likely.

We set the *control gene* to v_n, and find the optimal MCR intervention policy for a given Θ and Λ. For a given uncertainty class, and a given patient known to belong to this uncertainty class, our objective is to estimate $\pi_R^\bullet(U)$ and $S_R(U)$.

Population data. In this section, we evaluate how well $\hat{\Lambda}$ estimates Λ based on population data, and how estimates of our prognosis metrics based on $\hat{\Lambda}$ compare to optimal estimates based on Λ.

Consider an uncertainty class containing 20 networks. One of these networks corresponds to a single mutation, having probability 2/3, and the other 19 correspond to two mutations, each having probability 1/57. We observe population data consisting of $m = 10$, 50, 100, 500, or 1000 patients, each associated with a network drawn from the uncertainty class with the above probabilities. The data observed from each patient in the population data consists of either a contiguous block of $l_0 \equiv l_1 = \ldots = l_m$ time-series GAPs sampled from the true BNp without control, or l_0 independent GAPs sampled from the true SSD without control, where l_0 = 10, 50, 100, 500, or 1000. We assume the 20 networks in the uncertainty class are known, and impose a uniform Dirichlet prior on Λ with $\alpha(R) = 1$ for all $R \in \Theta$. We approximate $\hat\Lambda$ based on Eq. (15) with $k = 100$ Monte-Carlo iterations. This entire procedure is repeated over 100 iterations.

Figures 2A and 2B provide the average of

$$\| \Lambda - \hat\Lambda \| = \sqrt{\sum_{R\in\Theta} (\Lambda(R) - \hat\Lambda(R))^2}$$

over simulation iterations with respect to m and l_0. In Figure 2A, it was observed that the performance of $\hat\Lambda$ initially improves as we increase m, until it saturates around $m = 100$. This suggests that for a fixed observation length, l_0, beyond a point there is not much improvement in network inference when collecting more population data. In Figure 2B, observation length appears to be a very important factor in estimating Λ. In this example, it is better to have a large amount of data from a few patients, than a small amount of data from many patients. Also note that the estimation of Λ is generally not accurate, with $\| \Lambda - \hat\Lambda \|$ over 0.35 even with $m = 1000$ and $l_0 = 1000$.

In this uncertainty class, time-series data is better than independent data for estimating Λ, except when l_0 is very small. To understand this, Figure 3 shows the true network probabilities, Λ, and average estimated network probabilities, $\hat\Lambda$, over simulation iterations for $m = 1000$ and various l_0. Time-series data of length $l_0 = 1000$ appears to estimate the probability of most networks correctly, but some of the mass of network 1 has been spread to networks 3 and 18, indicating that networks 1, 3, and 18 are indistinguishable with this kind of data. Increasing l_0 to 10,000 improves the estimation of Λ, with only networks 1 and 18 being slightly difficult to distinguish. Independent data of length $l_0 = 1000$ appears to be unable to distinguish between networks 1, 3, 4, 12, 18, and 20, and increasing l_0 to 10,000 helps, but networks 1, 3, and 18 still cannot be distinguished. This *identifiability* problem may be caused by distinct networks in the uncertainty class having similar SSDs without control, thus data drawn from these networks come from similar distributions. This suggests

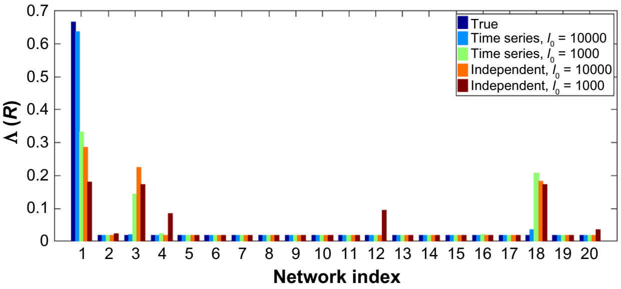

Figure 3. True network probabilities, Λ, for an uncertainty class with 20 networks, and various estimates, $\hat\Lambda$, for the same uncertainty class derived from $m = 1000$ simulated population data and averaged over 100 simulation iterations.

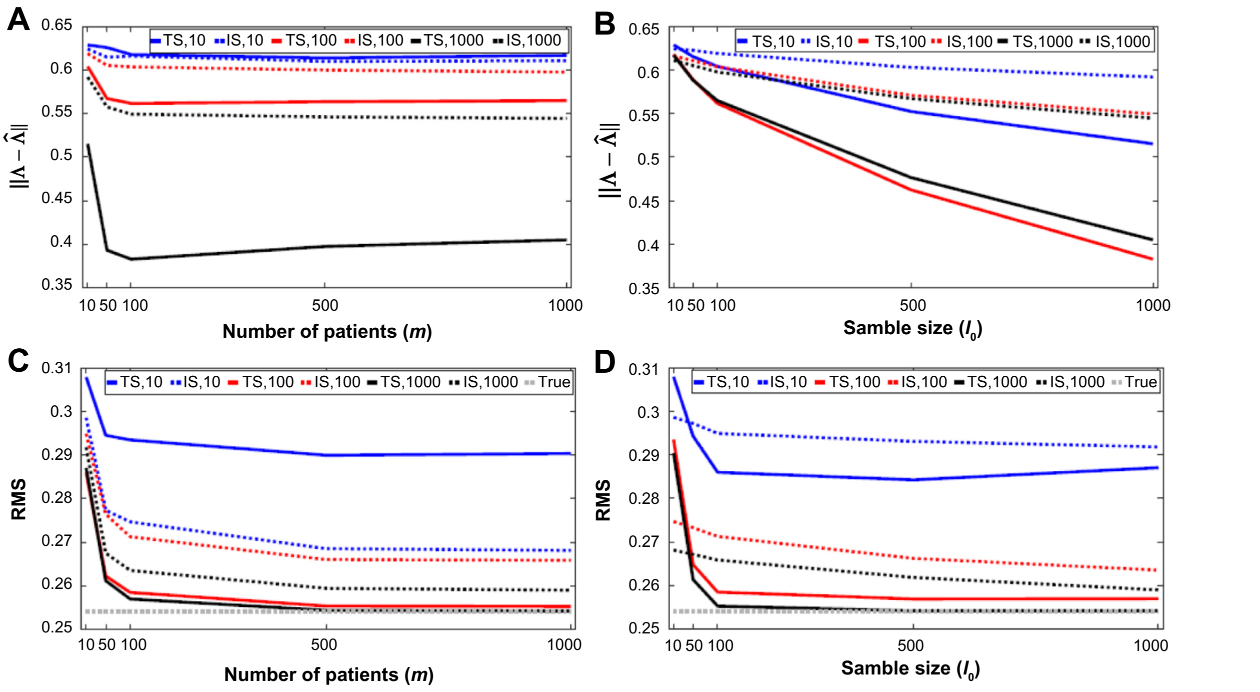

Figure 2. Estimating Λ and S_R (U) with population data under an uncertainty class that favors time-series data. TS indicates time series, IS indicates independent data: (**A**) the average of $\| \Lambda - \hat\Lambda \|$ over simulation iterations with respect to m for l_0 = 10, 100, and 1000, (**B**) the average of $\| \Lambda - \hat\Lambda \|$ over simulation iterations with respect to l_0 for m = 10, 100, and 1000, (**C**) the empirical RMS of $\hat{S}$ (U) with respect to m for l_0 = 10, 100, and 1000, and (**D**) the empirical RMS of $\hat{S}(U)$ with respect to l_0 for m = 10, 100, and 1000.

a tradeoff: inference with time-series data requires more data points to estimate a large TPM, but may resolve identifiability problems between networks with similar SSDs. Note that under small l_0, independent points may outperform because time-series points are correlated and thus contain redundancy. A similar phenomenon was observed in Ref. 62, where the authors investigated prediction under independent and correlated samples and concluded that a predictor trained with correlated points can perform either better or worse than with independent points depending on the settings.

Figures 2C and 2D provide the empirical RMS of $\hat{S}(U)$, found as the root of the average of

$$\sum_{R\in\Theta}\left(\hat{S}(U)-S_R(U)\right)^2\Lambda(R)$$

over simulation iterations, with respect to m and l_0. Note that $\hat{S}(U)$ is computed without knowledge of Λ, while the empirical RMS uses Λ to evaluate the accuracy of $\hat{S}(U)$. The performance of $\hat{S}(U)$ appears to saturate when we fix m or l_0, suggesting that both must increase for performance to improve. We nearly achieve the RMS of optimal prediction based on the true Λ without patient data, indicated by a horizontal dotted line, with time-series data and m and l_0 at least 100. Although it may not be possible to estimate Λ accurately for a given sample size, an inaccurate estimate of Λ may still carry information sufficient for prognosis prediction. To illustrate why, recall that networks 1, 3, and 18 are indistinguishable with time-series data of length $l_0 = 1000$, in the sense that the true mass of Λ at network 1 is spread to networks 3 and 18 in $\hat{\Lambda}$. As long as these networks have similar shift (and other prognosis metrics), the estimated shift computed in Eq. (9) is accurate. In Figure 2D, it was also observed that for all m, the performance of time-series and independent data cross at a higher threshold than in Figure 2B, at around $l_0 = 50$. Thus, the space of settings for which independent data outperform time-series data is larger for prognosis prediction than network inference.

We repeat this with a second uncertainty class having 230 networks and $k = 50$. In Figure 4A, as earlier, when l_0 is small, increasing m is not helpful, and independent data outperforms time-series data when estimating Λ. However, when l_0 is large, increasing m appears more helpful, and time-series data appears to have less of an advantage. When inferring prognosis in Figures 4C and 4D, independent data is better than time-series data, and RMS saturates around $l_0 = 100$ with independent data for all m. This is an example where independent data may be superior to time-series data, and there is not much improvement in prognosis prediction when collecting more than $l_0 = 100$ points from each patient for any m.

Patient data. In this section, we focus on estimates of prognosis metrics based on patient data. For each uncertainty class, we draw 100 time-series and 100 independent population datasets, with $m = 1000$ and $l_0 = 100$. All networks

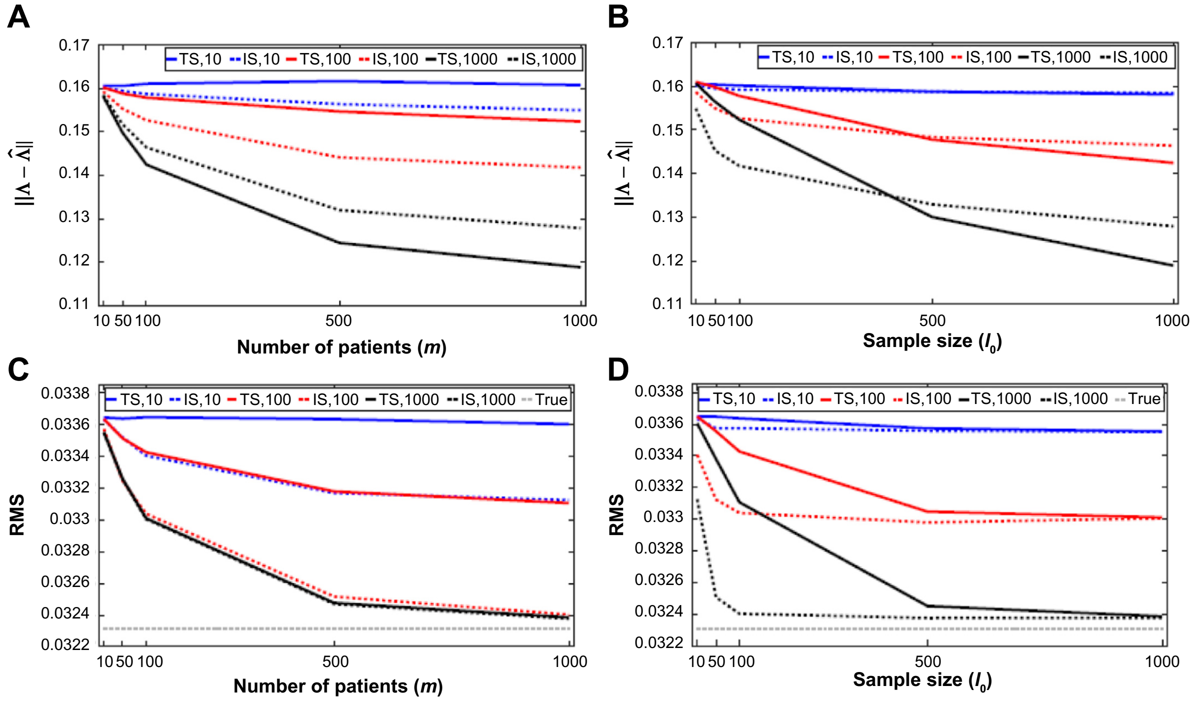

Figure 4. Estimating Λ and $S_R(U)$ with population data under an uncertainty class that favors independent data. TS indicates time series, IS indicates independent data: (**A**) the average of $\|\Lambda - \hat{\Lambda}\|$ over simulation iterations with respect to m for $l_0 = 10$, 100, and 1000, (**B**) the average of $\|\Lambda - \hat{\Lambda}\|$ over simulation iterations with respect to l_0 for $m = 10$, 100, and 1000, (**C**) the empirical RMS of $\hat{S}(U)$ with respect to m for $l_0 = 10$, 100, and 1000, and (**D**) the empirical RMS of $\hat{S}(U)$ with respect to l_0 for $m = 10$, 100, and 1000.

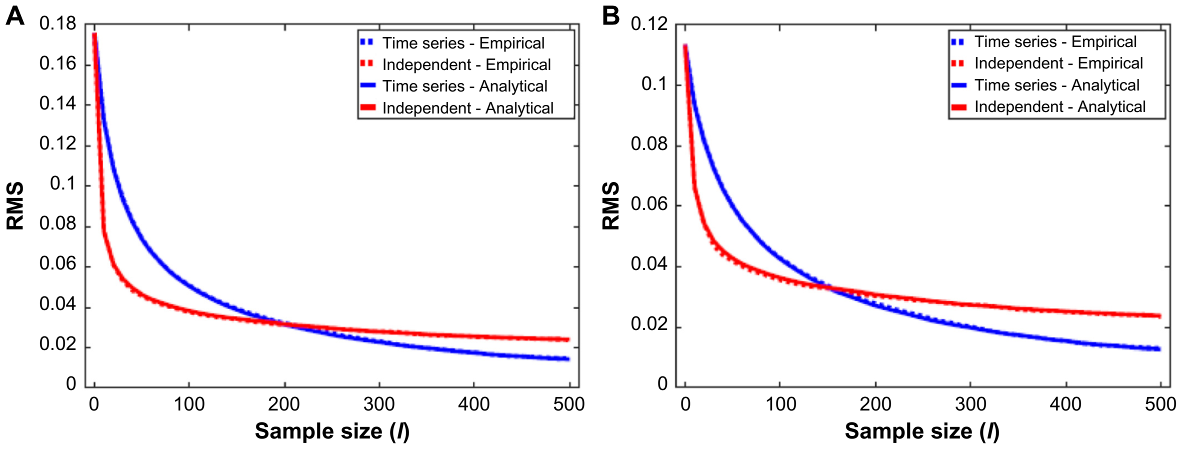

Figure 5. Analytical and empirical RMS with respect to *l* for independent and time-series data under the true Λ: (**A**) estimated undesirable mass with control and (**B**) estimated shift.

are drawn according to the true network probabilities, Λ. For each population dataset, we find $\hat{\Lambda}$ with $k = 100$. We draw 1000 time-series and 1000 independent patient datasets. Each patient dataset is associated with a network drawn from the true network probabilities, Λ. Assuming the true Λ is known, for each patient dataset of the independent type, we find $\hat{\pi}^\bullet(U)$, the analytical MSE of $\hat{\pi}^\bullet(U)$, $\hat{S}(U)$ and the analytical MSE of $\hat{S}(U)$ for different sizes of the data, *l*, constructed by starting with no data, and incrementally adding blocks of 10 independent points until we reach $l = 500$ points. For each time-series patient dataset, we incrementally add independent blocks of 10 time-series points until we reach $l = 500$ points. Prognosis predictions are also found for all combinations of independent population data (for each we effectively substitute $\hat{\Lambda}$ for the true network probabilities) and independent patient data, as well as all combinations of time-series population and time-series patient data. This entire procedure is repeated over 100 uncertainty classes.

In Figure 5A, the analytical RMS of $\hat{\pi}^\bullet(U)$, found as the root of the average of MSE $[\hat{\pi}^\bullet(U)]$ from Eq. (8) over all uncertainty classes and iterations, and an empirical RMS of $\hat{\pi}^\bullet(U)$, found as the root of the average of $\left(\hat{\pi}^\bullet(U) - \pi_R^\bullet(U)\right)^2$ over all uncertainty classes and iterations, were graphed with respect to *l* for independent and time-series data assuming the true Λ is known. Figure 5B provides analogous graphs for the analytical and empirical RMS of $\hat{S}(U)$. Analytical and empirical RMS curves coincide, as they must when Λ is known. Independent data outperforms time series when the data size is small with $l < 150$, and time series outperforms when the data size is large with $l > 200$.

Figure 6 shows scatter plots of true prognosis metrics (vertical axis) versus their estimates (horizontal axis) over 100 uncertainty classes and known Λ with no data, independent data of length $l = 500$, and time-series data of length $l = 500$. Correlation coefficients for each scatter plot are provided in the captions. With no data, each point in the scatter plots represents one network in one uncertainty class. Since all networks in a given uncertainty class are assigned the same estimated shift, we observe a vertical "stripe" for each uncertainty class. With data, each point represents one of 1000 iterations for an uncertainty class. In all cases, the linear regression line, shown as a solid black line, coincides with the ideal 45° dashed red line. A plot of correlation coefficients with respect to data size, *l*, is provided in Figure 7. As in Figure 5, independent data outperforms time-series data when $l < 150$, and time-series outperforms when $l > 200$.

Next consider a fixed uncertainty class with 146 networks. Figure 8 shows the analytical and empirical RMS of estimated prognosis metrics with respect to *l* for this uncertainty class, under independent and time-series data. Values based on the true Λ are averaged over 1000 replications of patient data, and values based on estimated Λ are averaged over all combinations of 100 replications of population data with $m = 1000$ and $l_0 = 100$, and 1000 replications of patient data. When Λ must be estimated, RMS increases. Furthermore, independent data results in lower RMS than time-series data in this uncertainty class.

Figure 9 shows correlation graphs across all patients in this uncertainty class, for true versus estimated prognosis metrics based on the true Λ, true versus estimated prognosis metrics based on estimated Λ, and estimated prognosis metrics based on the true versus estimated Λ, all with respect to *l* under independent and time-series data. As in the RMS curves, independent data results in higher correlation than time-series data for this example. The correlation of estimated prognosis metrics based on the true versus estimated Λ is nearly 1 for almost all sample sizes considered.

Figure 10 illustrates the analytical RMS of our prognosis metrics for each patient with known Λ. Each row represents a patient, which have been grouped so that patients with the same underlying network are next to each other, each column represents a patient data size, *l*, and a darker color indicates a

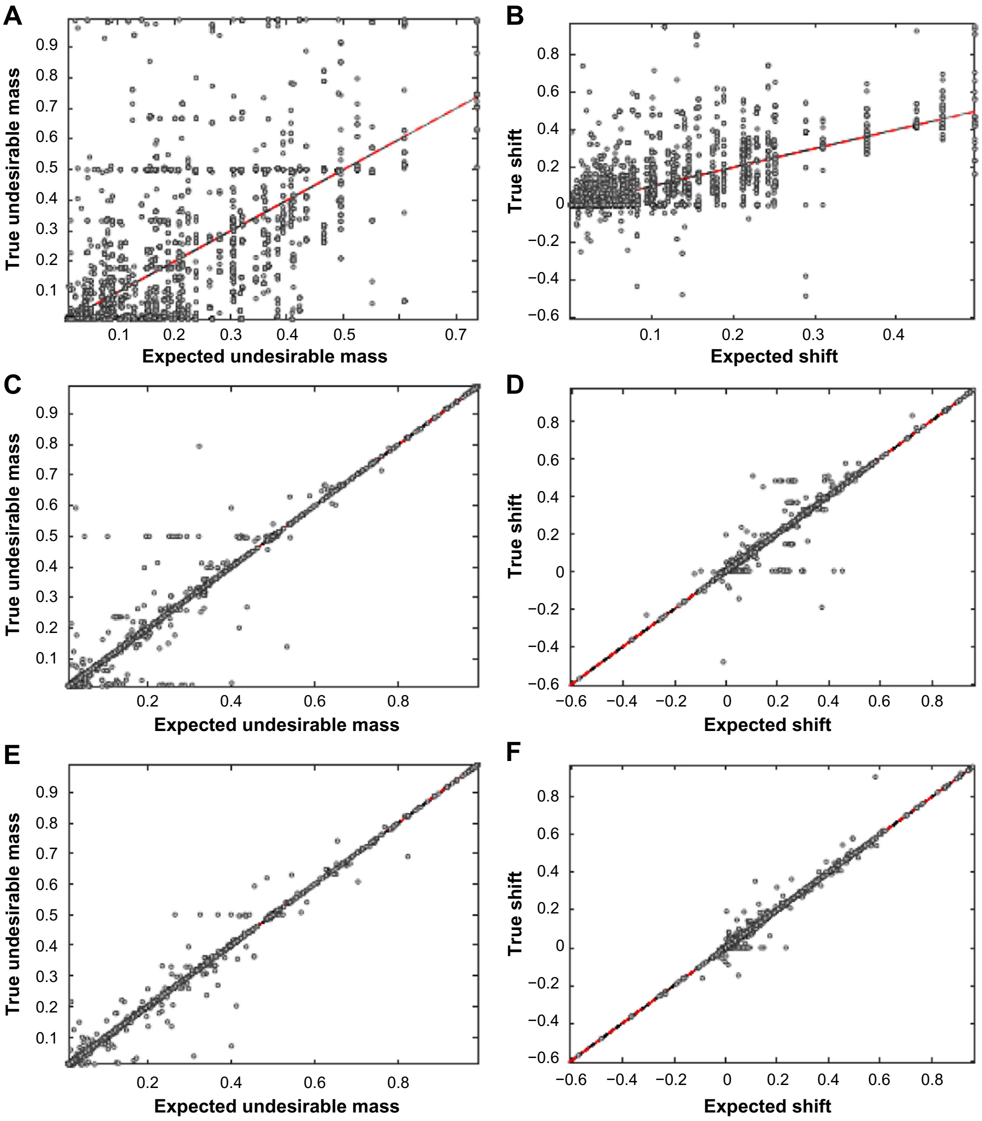

Figure 6. Scatter plots of true prognosis metrics (vertical axis) versus their estimates (horizontal axis) over 100 uncertainty classes and known Λ: (**A**) undesirable mass with control, no data ($\rho = 0.6704$), (**B**) shift, no data ($\rho = 0.6881$), (**C**) undesirable mass with control, independent data with $l = 500$ ($\rho = 0.9950$), (**D**) shift, independent data with $l = 500$ ($\rho = 0.9882$), (**E**) undesirable mass with control, time-series data with $l = 500$ ($\rho = 0.9981$), and (**F**) shift, time-series data with $l = 500$ ($\rho = 0.9963$).

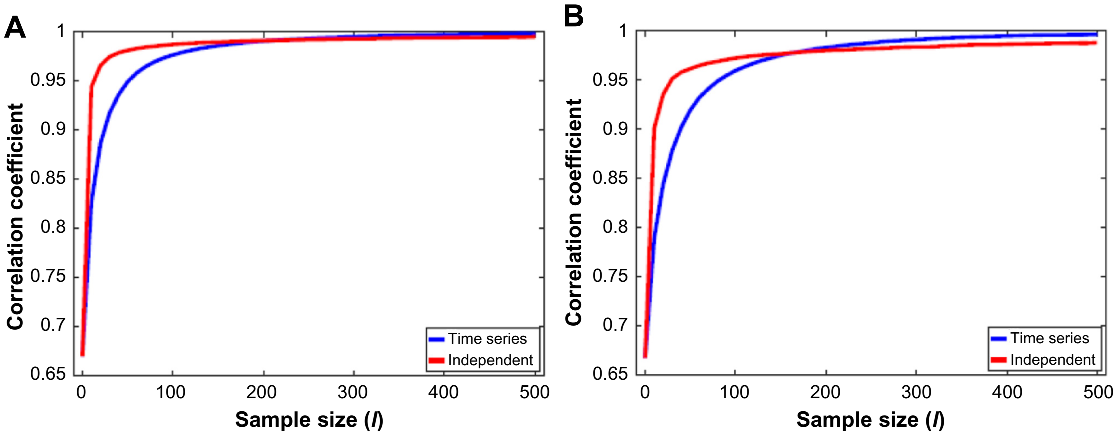

Figure 7. Correlation of true versus estimated prognosis metrics over 100 uncertainty classes and 1000 patient iterations per uncertainty class. Independent and time-series data are shown for known Λ: (**A**) undesirable mass with control and (**B**) shift.

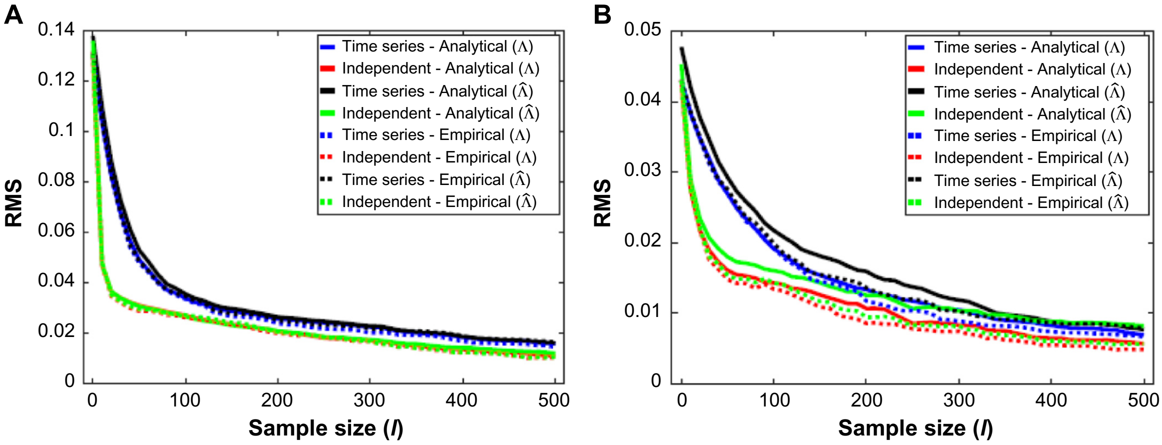

Figure 8. Analytical and empirical RMS with respect to *l* for a fixed uncertainty class, independent and time-series data: (**A**) undesirable mass with control and (**B**) shift.

lower sample-conditioned RMS. With no data (the leftmost column), all patients have the same analytical RMS. Estimated prognosis under certain networks tend to converge at a faster rate, whereas others tend to require more data. Also note the RMS of undesirable mass and shift have similar trends, but there are some networks for which one converges faster than the other.

In Figure 11, we present an example of the convergence of estimated prognosis metrics and their analytical MSE to the true prognosis metrics for a particular patient dataset in this uncertainty class. The horizontal axis represents undesirable mass with control, and the vertical axis represents shift. The true network prognosis is indicated by a red dot. With no data, we obtain the mean, variance, and covariance of these prognosis metrics from our MMSE estimates, the analytical MSE of these estimators, and the analytical covariance in Eq. (11). Our initial knowledge of the prognosis metrics is illustrated in the figure with a large ellipse, which is effectively the unit-standard-deviation level curve of a Gaussian distribution with the appropriate mean and covariance matrix. Smaller ellipses represent the same procedure with increasing data size, up to $l = 200$ in steps of 20. Knowledge converges to a near certainty around the true prognosis.

Cell-cycle Network Simulation

We use a mammalian cell-cycle network to demonstrate the application of our approach to a biologically motivated example. This network models normal biological processes in mammalian cells during the cell cycle, whose disruption can lead to uncontrolled cell proliferation.[63] The network has 10 genes (or proteins), CycD, Rb, p27, E2F, CycE, CycA, Cdc20, Cdhl, UbcHIO, and CycB, where regulatory relationships in R^H are illustrated in Figure 12. Three key elements in this network are Cyclin D (CycD), retinoblastoma (Rb), and p27. Extracellular signals, under normal conditions, coordinate cell division with overall growth by controlling the activation of

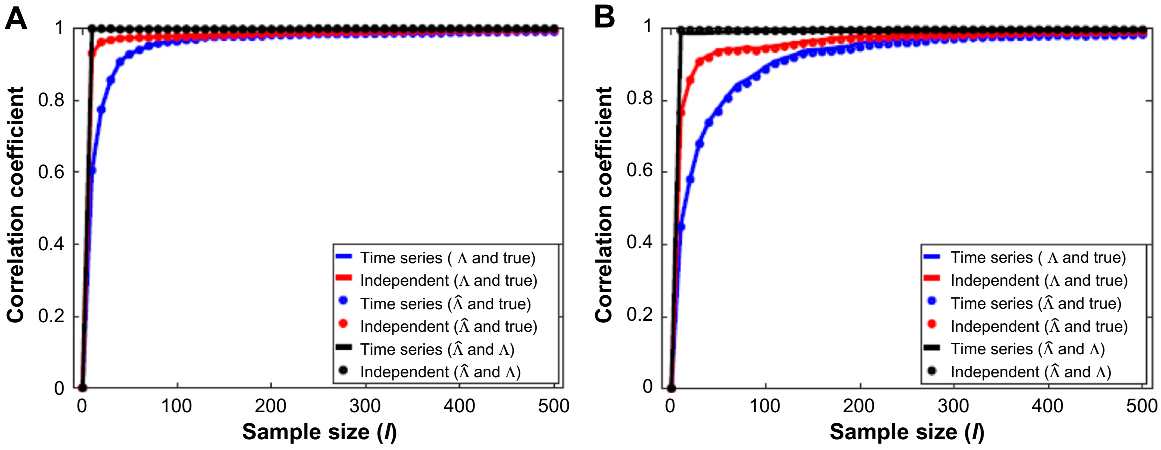

Figure 9. Correlation between true prognosis metrics, estimated prognosis metrics with known Λ, and estimated prognosis metrics with estimated Λ for a fixed uncertainty class and 1000 patient iterations. Independent and time-series data are shown: (**A**) undesirable mass with control and (**B**) shift.

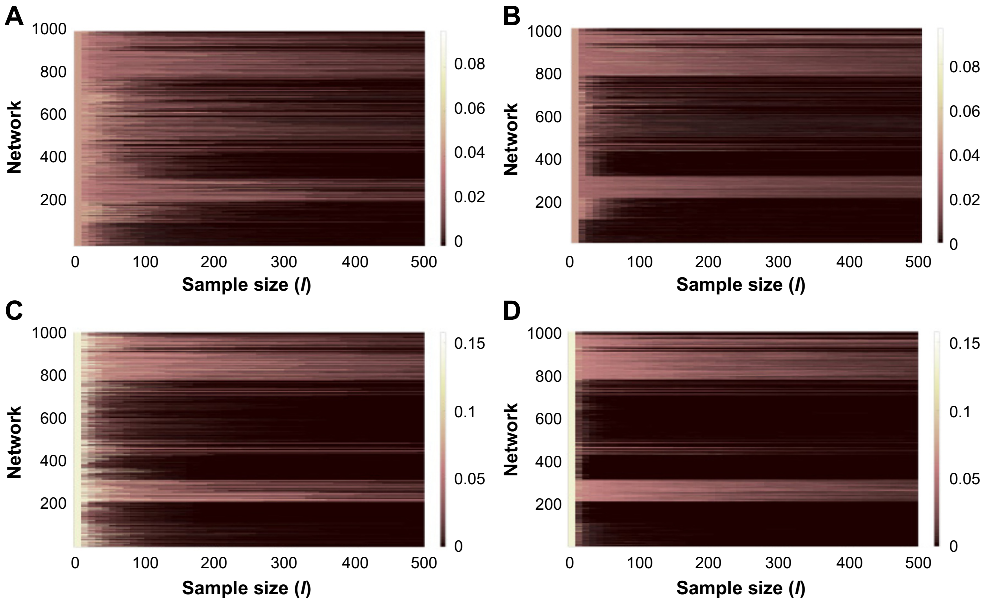

Figure 10. Analytical RMS of prognosis metrics for each patient in each uncertainty class with known Λ. Each row represents a patient, grouped by networks in a fixed uncertainty class, each column represents a patient data size, *l*, and color indicates analytical RMS for a prognosis metric: (**A**) shift with time-series data, (**B**) shift with independent data, (**C**) undesirable mass with control with time-series data, and (**D**) undesirable mass with control with independent data.

CycD. The tumor suppressor protein, Rb, plays a central role in the negative control of the cell cycle and in tumor progression.[64] Rb is active in the absence of certain cyclin-dependent kinases. When present, however, these cyclins inhibit Rb through phosphorylation. p27 is also active in the absence of the cyclins. An active p27 blocks the action of CycE or CycA, and hence, Rb can also be active even if the cyclins are present, resulting in a mechanism that stops uncontrolled cell division. On the other hand, simultaneous inactivation of CycD, Rd, and p27 may lead to cell proliferation, which is undesirable in the absence of any growth factor. This motivates us to define the set of *undesirable states* to correspond to the condition where CycD = Rd = p27 = 0.

Assuming the same majority-voting rule and setting p = 0.01, we construct a BNp for the healthy network. The number of networks in the uncertainty class grows rapidly as we compound mutations, thus we only include regulatory networks with one edge removed from R^H in our uncertainty class. We also exclude from this set those networks that have a lower undesirable steady-state mass than either the healthy network or the average undesirable mass of all networks in the set.

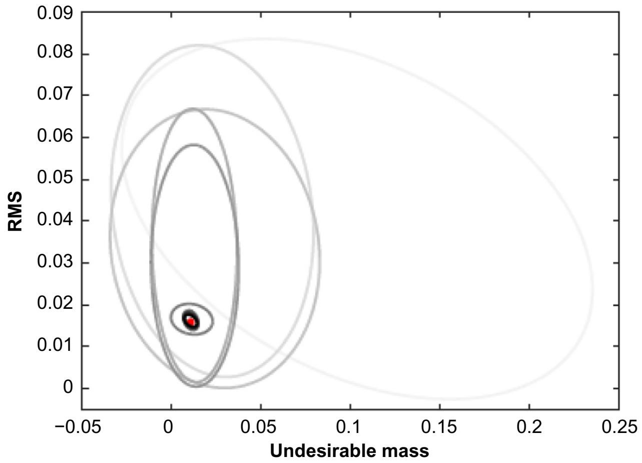

Figure 11. Example of convergence of estimated prognosis metrics and their analytical MSE to the true prognosis metrics. The intensity of color in each ellipse indicates the amount of data, where darker lines indicate more patient data.

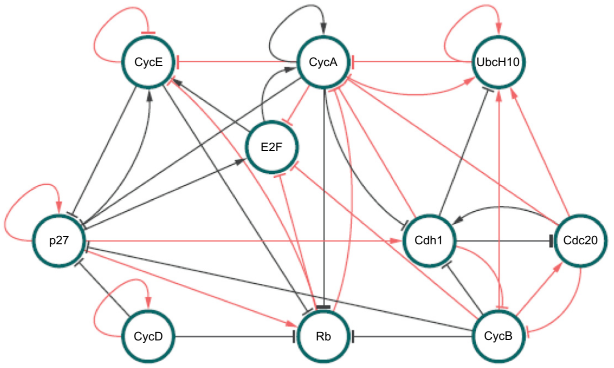

Figure 12. Regulatory graph for a normal mammalian cell cycle network. Arrows indicate activation, corresponding to a +1 in the regulatory matrix, and blunted arrows indicate inhibition, corresponding to a –1 in the regulatory matrix. Twenty-one mutations of the normal network are considered in the uncertainty class, each associated with the removal of a single red edge.

Table 1. $\| \Lambda - \hat{\Lambda} \|$ with time-series/independent data.

	$m = 10$	$m = 100$	$m = 1000$
$l_0 = 1$	0.2964/0.2947	0.3020/0.3034	0.2933/0.2940
$l_0 = 10$	0.2972/0.2930	0.2965/0.2848	0.2837/0.2737
$l_0 = 100$	0.2867/0.2805	0.2653/0.2435	0.2551/0.2328

The uncertainty class, Θ, resulting from this procedure contains 21 mutated networks, each found by removing a single red edge in Figure 12. We set the true probability distribution, Λ, based on gene alteration rates in breast cancer as reported by TCGA data retrieved from cBioPortal.[65,66] We find the MCR intervention policy, which maximally shifts the undesirable steady-state mass averaged across Θ, with E2F as the control gene. Our methodology can be implemented with any control gene, although in general this choice affects the particular intervention outcome, and thus the prognosis outcome. Simulation settings that are not specified are the same as in synthetic data simulations.

Table 1 provides $\| \Lambda - \hat{\Lambda} \|$ for various l_0 and m. In this particular uncertainty class, l_0 appears to play a more important role than m in inferring Λ, although $\| \Lambda - \hat{\Lambda} \|$ is larger than 0.25 in all cases. Figure 13 shows the mean of true and estimated prognosis metrics, and Figure 14 shows the analytical and empirical RMS of prognosis metrics. Results with both time-series and independent data, and both Λ known and $\hat{\Lambda}$ estimated with population data of size $l_0 = 1$ and $m = 1000$, are shown in each figure. In this example, independent data is superior to time-series data for the entire range of patient sample sizes shown. For instance, under independent data with only $l = 100$, the mean of estimated prognosis metrics have nearly converged to the mean of the true metrics. The RMS of independent data is about 0.01, while the RMS of time-series data for the same sample size is at about 0.02. Prognosis prediction with $\hat{\Lambda}$ is comparable to prediction with the true Λ, thus, the effect of increasing patient data size is much more profound than increasing population data size.

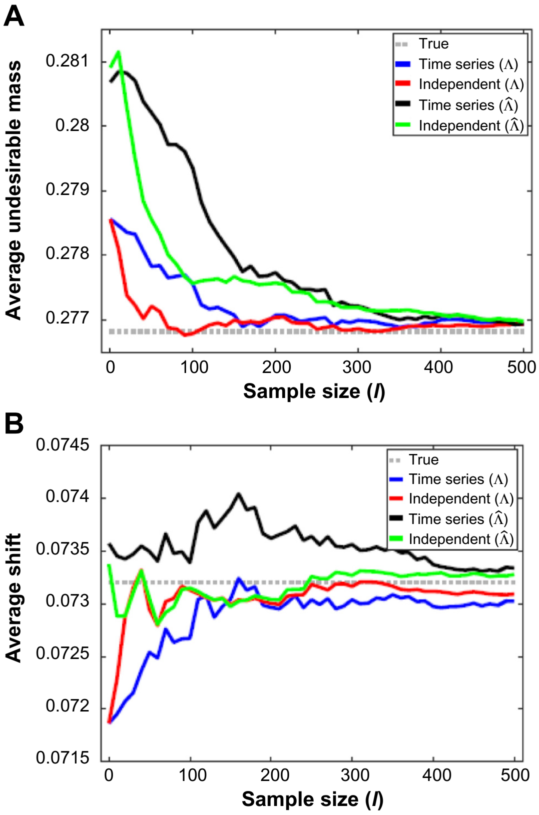

Figure 13. Mean of true and estimated and prognosis metrics with respect to l for the mammalian cell-cycle network, independent and time-series data: (**A**) undesirable mass with control and (**B**) shift.

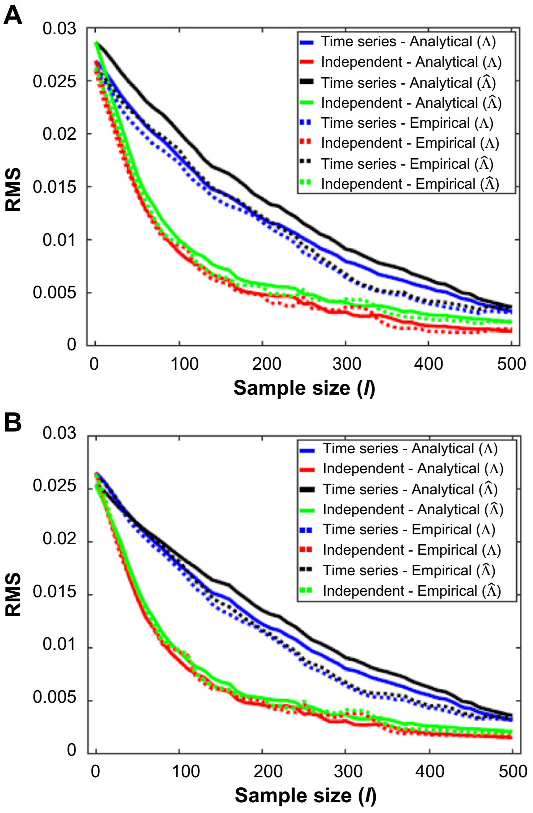

Figure 14. Analytical and empirical RMS with respect to l for the mammalian cell-cycle network, independent and time-series data: (**A**) undesirable mass with control and (**B**) shift.

Conclusion

We have outlined a general model-based framework to optimally predict prognosis. Our work begins with many optimistic and simplifying assumptions, for instance that regulatory networks are well modeled by BNps, and that a healthy BNp can be fully determined from available scientific knowledge. This reflects our belief that reliable cancer prognosis and prediction should begin with understanding and modeling healthy cell functioning (including accounting for incomplete knowledge of normal regulation), as suggested in Ref. 23. Once healthy gene regulation is understood, ie, once R^H is known in our modeling framework, population data is only helpful to inform about mutations and their frequencies, and as we have seen, prognosis prediction is not highly sensitive to the estimate of Λ. This is an intuitive result: cancer is a personalized disease and inference for given patient rests critically on what data is available from the patient of interest. If prior knowledge of gene regulation is not available or only partially available, then population data may serve an important role in discovering new scientific knowledge, which may be used to construct an uncertainty class of gene regulatory models.

Although the performance of prognosis prediction depends on many factors, including the networks constituting the uncertainty class and the individual patient's network, performance primarily depends on how much patient data is available, and the extent to which prognosis is identifiable with the available type of data. If one's interest is in fully inferring the patient network, typically time-series data is better than independent data, as it is possible that two networks in the uncertainty class have very similar SSDs, representing similar phenotypes, but very different TPMs, representing different regulatory connections. Even time-series data may not be sufficient for network inference, since it is possible for two networks to represent different mutations and yet have very similar dynamics. In this case, experiments specifically designed to identify mutations may be necessary to infer a network.

If one's interest is in inferring prognosis, we have seen that independent data performs remarkably well, particularly under small samples. This is partly because our metrics for prognosis, $\pi^{\bullet}_{R}(U)$ and $S_R(U)$, depend only on the undesirable mass of SSDs of the network with and without control. Independent data drawn without control reveals information about the SSD without control, and as long as networks with the same SSD without control also have similar undesirable mass with control, our metrics for prognosis can be identified with independent data. At the same time, small time-series data may carry less information due to correlations.

To bring this work to the bedside, several extensions can be explored, including: (1) more realistic regulatory network models, drug models, and prognosis metrics, (2) multiple healthy networks, (3) incomplete prior information about healthy or unhealthy networks and their parameters, (4) more realistic observation models for RNA-seq, protein mass spectrometry and other types of data acquisition technologies currently available to clinicians, (5) allowing the network to change or deteriorate over time, and (6) design of experiments or treatment regimens tailored to an individual patient. Our framework may also be extended to classify or infer other relevant metrics, for instance, disease subtype, mean survival time, drug response, or drug side effects.

Acknowledgments

The results published here are in part based upon data generated by The Cancer Genome Atlas (TCGA) established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at http://cancergenome.nih.gov.

Author Contributions

Conceived and designed the experiments: LAD, MRY. Analyzed the data: LAD, MRY. Wrote the first draft of the manuscript: LAD, MRY. Contributed to the writing of the manuscript: LAD, MRY. Agree with manuscript results and conclusions: LAD, MRY. Jointly developed the structure and arguments for the paper: LAD, MRY. Made critical revisions and approved final version: LAD, MRY. Both authors reviewed and approved of the final manuscript.

REFERENCES

1. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011;144:646–74.
2. McGuire WL. Breast cancer prognostic factors: evaluation guidelines. *J Natl Cancer Inst*. 1991;83:154–5.
3. Stephenson CF, Bridge JA, Sandberg AA. Cytogenetic and pathologic aspects of Ewing's sarcoma and neuroectodermal tumors. *Hum Pathol*. 1992;23:1270–7.
4. Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 2000;403:503–11.
5. Veer LJ, Dai H, Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;415:530–6.
6. Glinsky GV, Glinskii AB, Stephenson AJ, Hoffman RM, Gerald WL. Gene expression profiling predicts clinical outcome of prostate cancer. *J Clin Invest*. 2004;113:913–23.
7. Paik S, Tang G, Shak S, et al. Gene expression and benefit of chemotherapy in women with node-negative estrogen receptor-positive breast cancer. *J Clin Oncol*. 2006;24:3726–34.
8. Ray T. FDA's woodcock says personalized drug development entering 'long slog' phase. *Pharmacogenomics Reporter*. Available at: http://www.genomeweb.com/mdx/fdas-woodcock-says-personalized-drug-development-entering-long-slog-phase. Accessed October 26, 2011.
9. Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*. 2005;21:171–8.
10. Ioannidis JPA. Why most published research findings are false. *PLoS Med*. 2005;2:e124.
11. Zhang M, Yao C, Guo Z, et al. Apparently low reproducibility of true differential expression discoveries in microarray studies. *Bioinformatics*. 2008;24:2057–63.
12. Boulesteix AL. Over-optimism in bioinformatics research. *Bioinformatics*. 2010;26: 437–9.
13. Yousefi MR, Dougherty ER. Performance reproducibility index for classification. *Bioinformatics*. 2012;28:2824–33.
14. Soneson C, Gerster S, Delorenzi M. Batch effect confounding leads to strong bias in performance estimates obtained by cross-validation. *PLoS One*. 2014;9: e100335.
15. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. 2008;18:1509–17.
16. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci U S A*. 2006;103:5923–8.
17. Draghici S, Khatri P, Eklund AC, Szallasi Z. Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet*. 2006;22:101–9.
18. Sol A, Balling R, Hood L, Galas D. Diseases as network perturbations. *Curr Opin Biotechnol*. 2010;21:566–71.
19. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. *Nat Med*. 2004;10:789–99.
20. Ideker T, Dutkowski J, Hood L. Boosting signal-to-noise in complex biology: prior knowledge is power. *Cell*. 2011;144:860–3.
21. Huang S, Kauffman SA. How to escape the cancer attractor: rationale and limitations of multi-target drugs. *Semin Cancer Biol*. 2013;23:270–8.
22. Schork NJ. Personalized medicine: time for one-person trials. *Nature*. 2015;520: 609–11.
23. Krogan N, Lippman S, Agard D, Ashworth A, Ideker T. The cancer cell map initiative: defining the hallmark networks of cancer. *Mol Cell*. 2015;58:690–8.
24. Loscalzo J, Barabasi AL. Systems biology and the future of medicine. *Wiley Interdiscip Rev Syst Biol Med*. 2011;3:619–27.

25. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102:15545–50.
26. Yu JX, Sieuwerts AM, Zhang Y, et al. Pathway analysis of gene signatures predicting metastasis of node-negative primary breast cancer. *BMC Cancer*. 2007;7:182.
27. Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. *Mol Syst Biol*. 2007;3:140.
28. Lee E, Chuang HY, Kim JW, Ideker T, Lee D. Inferring pathway activity toward precise disease classification. *PLoS Comput Biol*. 2008;4:e1000217.
29. Su J, Yoon BJ, Dougherty ER. Accurate and reliable cancer classification based on probabilistic inference of pathway activity. *PLoS One*. 2009;4:e8161.
30. Dao P, Wang K, Collins C, Ester M, Lapuk A, Sahinalp SC. Optimally discriminative subnetwork markers predict response to chemotherapy. *Bioinformatics*. 2011;27:i205–13.
31. Winter C, Kristiansen G, Kersting S, et al. Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput Biol*. 2012;8:e1002511.
32. Roy J, Winter C, Isik Z, Schroeder M. Network information improves cancer outcome prediction. *Brief Bioinform*. 2012;15:612–25.
33. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. *Nat Methods*. 2013;10:1108–15.
34. Allahyar A, Ridder JD. FERAL: network-based classifier with application to breast cancer outcome prediction. *Bioinformatics*. 2015;31:i311–9.
35. Staiger C, Cadot S, Kooter R, et al. A critical evaluation of network and pathway-based classifiers for outcome prediction in breast cancer. *PLoS One*. 2012;7:e34796.
36. Cun Y, Frhlich HF. Prognostic gene signatures for patient stratification in breast cancer – accuracy, stability and interpretability of gene selection approaches using prior knowledge on protein-protein interactions. *BMC Bioinformatics*. 2012;13:69.
37. Staiger C, Cadot S, Gyrffy B, Wessels LF, Klau GW. Current composite-feature classification methods do not outperform simple single-genes classifiers in breast cancer prognosis. *Front Genet*. 2013;4:289.
38. Bilal E, Dutkowski J, Guinney J, et al. Improving breast cancer survival analysis through competition-based multidimensional modeling. *PLoS Comput Biol*. 2013;9:e1003047.
39. Shahrokh Esfahani M, Knight J, Zollanvari A, Yoon BJ, Dougherty ER. Classifier design given an uncertainty class of feature distributions via regularized maximum likelihood and the incorporation of biological pathway knowledge in steady-state phenotype classification. *Pattern Recognit*. 2013;46:2783–97.
40. Shahrokh Esfahani M, Dougherty ER. An optimization-based framework for the transformation of incomplete biological knowledge into a probabilistic structure and its application to the utilization of gene/protein signaling pathways in discrete phenotype classification. *IEEE/ACM Trans Comput Biol Bioinform*. 2015;12:1304–21.
41. Shahrokh Esfahani M, Dougherty ER. Incorporation of biological pathway knowledge in the construction of priors for optimal bayesian classification. *IEEE/ACM Trans Comput Biol Bioinform*. 2014;11:202–18.
42. McAdams HH, Arkin A. Stochastic mechanisms in gene expression. *Proc Natl Acad Sci U S A*. 1997;94:814–9.
43. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. *J Comput Biol*. 2000;7:601–20.
44. Chen KC, Wang TY, Tseng HH, Huang CYF, Kao CY. A stochastic differential equation model for quantifying transcriptional regulatory network in *Saccharomyces cerevisiae*. *Bioinformatics*. 2005;21:2883–90.
45. Knight J, Datta A, Dougherty ER. Generating stochastic gene regulatory networks consistent with pathway information and steady-state behavior. *IEEE Trans Biomed Eng*. 2012;59:1701–10.
46. Murrugarra D, Veliz-Cuba A, Aguilar B, Arat S, Laubenbacher R. Modeling stochasticity and variability in gene regulatory networks. *EURASIP J Bioinform Syst Biol*. 2012;2012:5.
47. Csermely P, Korcsmros T, Kiss HJM, London G, Nussinov R. Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacol Ther*. 2013;138:333–408.
48. Kauffman SA. *The Origins of Order: Self Organization and Selection in Evolution*. USA: Oxford University Press; 1993.
49. Shmulevich I, Dougherty ER, Kim S, Zhang W. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*. 2002;18:261–74.
50. Mukherjee S, Speed TP. Network inference using informative priors. *Proc Natl Acad Sci U S A*. 2008;105:14313–8.
51. Pal R, Datta A, Dougherty ER. Bayesian robustness in the control of gene regulatory networks. *IEEE Trans Signal Process*. 2009;57:3667–78.
52. Yoon BJ, Qian X, Dougherty ER. Quantifying the objective cost of uncertainty in complex dynamical systems. *IEEE Trans Signal Process*. 2013;61:2256–66.
53. Yousefi MR, Dougherty ER. A comparison study of optimal and suboptimal intervention policies for gene regulatory networks in the presence of uncertainty. *EURASIP J Bioinform Syst Biol*. 2014;2014:1–13.
54. Dehghannasiri R, Yoon BJ, Dougherty ER. Optimal experimental design for gene regulatory networks in the presence of uncertainty. *IEEE/ACM Trans Comput Biol Bioinform*. 2015;12:938–50.
55. Dalton LA, Yousefi MR. Optimal Bayesian cancer prognosis with model-constrained robust intervention. In: 2014 *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. Atlanta, GA. IEEE, NY; 2014:1382–5.
56. Yousefi MR, Dalton LA. Optimal cancer prognosis under network uncertainty. *EURASIP J Bioinform Syst Biol*. 2015;2015:1.
57. Yousefi MR, Ivanov I. Optimal control of gene regulatory networks with uncertain intervention effects. In: 2013 *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. Austin, TX. IEEE, NY; 2013:105–8.
58. Pal R, Datta A, Dougherty ER. Robust intervention in probabilistic Boolean networks. *IEEE Trans Signal Process*. 2008;56:1280–94.
59. Denic S, Vasic B, Charalambous C, Palanivelu R. Robust control of uncertain context-sensitive probabilistic Boolean networks. *IET Syst Biol*. 2009;3:279–95.
60. Derman C. *Finite State Markovian Decision Processes*. New York: Academic Press; 1970.
61. Yousefl MR, Dougherty ER. Intervention in gene regulatory networks with maximal phenotype alteration. *Bioinformatics*. 2013;29:1758–67.
62. Zollanvari A, Hua J, Dougherty ER. Analytical study of performance of linear discriminant analysis in stochastic settings. *Pattern Recognit*. 2013;46:3017–29.
63. Faure A, Naldi A, Chaouiya C, Thieffry D. Dynamical analysis of a generic Boolean model for the control of the mammalian cell cycle. *Bioinformatics*. 2006;22:e124–31.
64. Giacinti C, Giordano A. RB and cell cycle progression. *Oncogene*. 2006;25:5220–7.
65. Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*. 2012;2:401–4.
66. Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal*. 2013;6:11.