SCIENTIFIC REPORTS

OPEN

Uncovering hidden disease patterns by simulating clinical diagnostic processes

Received: 31 May 2017
Accepted: 24 January 2018
Published online: 05 February 2018

Abolfazl Ramezanpour[1,2] & Alireza Mashaghi[1]

Choosing a sequence of observations (often with stochastic outcomes) which maximizes the information gain from a system of interacting variables is essential for a wide range of problems in science and technology, such as clinical diagnostic problems. Here, we use a probabilistic model of diseases and signs/symptoms to simulate the effects of medical decisions on the quality of diagnosis by maximizing an appropriate objective function of the medical observations. The study provides a systematic way of proposing new medical tests, considering the significance of diseases and cost of the suggested observations. The efficacy of methods and role of the objective functions as well as initial signs/symptoms are examined by numerical simulations of the diagnostic process by exhaustive or Monte Carlo sampling algorithms.

Clinical diagnosis is typically made through a process that starts with identifying initial findings and noting the past medical history of the patient and ends with a diagnosis or unresolved differential diagnoses[1,2]. In practice, the sequence of steps one clinician follows may be very different from those taken by another clinician, and the same clinician may approach the problem differently in two nearly identical cases[3]. This variability in diagnostic approach has a complex source and is rooted in the limited and varied extent of knowledge of the clinicians, stochasticity of the decision-making process, and lack of solid risk and cost assessment strategies among others.

Since the classic paper by Ledley and Lusted[4] where they first detailed on how logic and probabilistic reasoning form the backbone of medical reasoning, there has been much progress in the development of diagnostic decision support systems (DDSS)[5–11]. In recent decades, due to limited availability of appropriate clinical data, there has been growing interest in developing heuristic formal and rigorous mathematical models. These studies covered a wide range of approaches from simple Bayesian models to Bayesian belief networks and neural networks[12–18].

Here, using simple and rigorous models, we look for determinants of the efficiency of a diagnostic approach, i.e. choice of a sequence of events that leads to a diagnosis. The study involves concepts and tools of machine learning and inference, as well as stochastic optimization, to deal with the model construction and the stochastic nature of the problem[19–22]. In ref.[23] we used techniques from statistical physics of disordered systems to study this problem with more emphasis on the role of the interaction graph of signs (hereafter, we refer to symptoms or signs as "signs" for simplicity) and diseases in the quality of diagnosis[24–28]. Our models are indeed natural generalizations of the simpler probabilistic models studied in previous works[13–15], which usually assume that only one disease is behind the findings (exclusive diseases assumption) or the diseases act independently on the signs (causal independence assumption). Moreover, for computational simplicity, it is usually assumed that there is no disease-disease and sign-sign interactions. We showed that such interactions can significantly improve the accuracy of diagnosis without resorting to the exclusive diseases or the causal independence assumption. In this paper, we extend our previous study by introducing new performance measures and optimization algorithms with more focus on the role of the objective function and initial number of observations in the performance of the diagnostics algorithms.

Given a model of disease and sign variables, we aim to propose an optimal sequence of medical tests maximizing an appropriate objective function of the observations (Fig. 1). Here, besides the nature of the model, the structure of the objective function and the initial number and quality of medical tests play a significant role. A reasonable objective function for these kind of problems is provided by the maximum value of the disease likelihood[29]. To reduce the diagnosis time and the mortality and morbidity of diseases, we propose an objective function which gives more weight to the more polarizing observations and dangerous diseases. We see how the initial

[1]Leiden Academic Centre for Drug Research, Faculty of Science, Leiden University, Leiden, The Netherlands. [2]Department of Physics, University of Neyshabur, Neyshabur, Iran. Correspondence and requests for materials should be addressed to A.M. (email: a.mashaghi.tabari@lacdr.leidenuniv.nl)

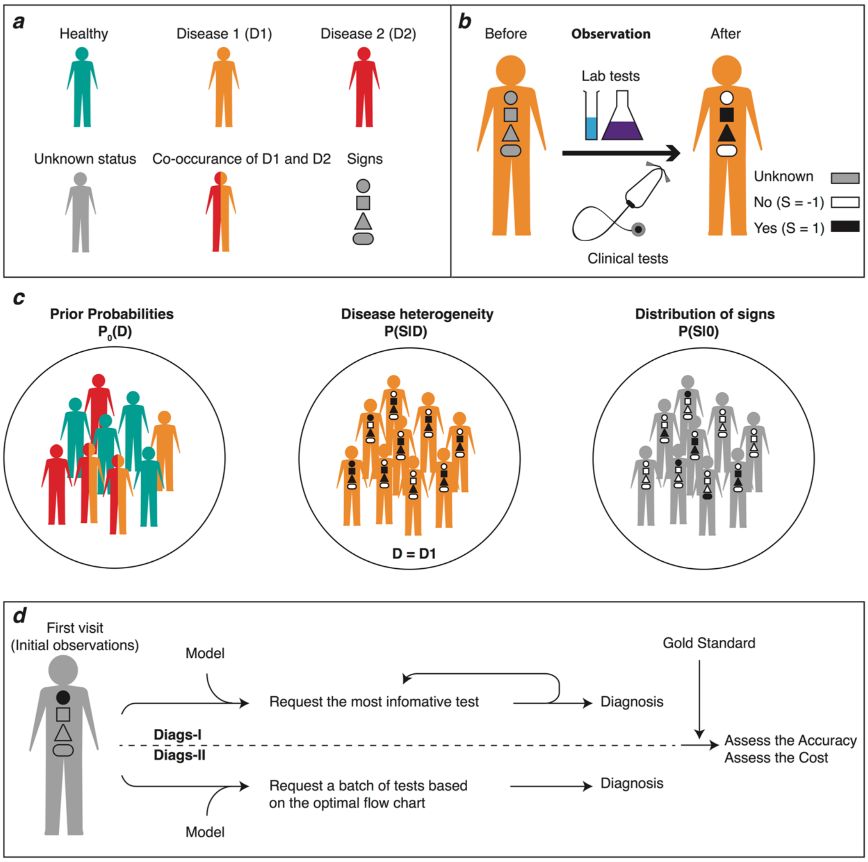

Figure 1. An illustration of the model definitions and the diagnostic processes. (**a**) A patient is represented with a disease pattern **D** (with **0** for the healthy state) and signs **S**. (**b**) A medical test changes an unobserved sign to an observed one with values ±1. (**c**) The probabilistic model is defined with the prior disease probabilities $P_0(\mathbf{D})$ and the conditional sign probabilities $P(\mathbf{S}|\mathbf{D})$. The leak probability $P(\mathbf{S}|\mathbf{0})$ takes into account the effects of unknown or ignored diseases. (**d**) The two diagnostic procedures (Diags-I and Diags-II) start from the same initial findings, but differ in the way the new observations are decided. In Diags-I, the true value of an observed sign is revealed by a medical test before going to the next observation. In Diags-II, the whole process is simulated with the sign values that are inferred from the probabilistic model.

number of observations and the cost of medical tests in the objective function affect the diagnosis performances in numerical simulations of the models. We also devise an approximate optimization algorithm based on the Monte Carlo sampling to construct an optimal sequence of medical tests for observation.

Main definitions and problem statement

Models. Consider a set of N_D binary variables $\mathbf{D} = \{D_a = 0, 1: a = 1, \ldots, N_D\}$, where $D_a = 0,1$ shows the absence or presence of disease a. We have another set of N_S binary variables $\mathbf{S} = \{S_i = \pm 1: i = 1, \ldots, N_S\}$ to show the values of sign variables (clinical and laboratory findings). We take W_a for the weight or importance of disease a, and C_i for the cost of observing sign i. In the following, the weights $W_a \in (0, 1)$ and costs $C_i \in (0, 1)$ are independent and identically distributed random variables with a uniform probability distribution. The joint probability distribution of the sign and disease variables (i.e., the model) is identified by $P(\mathbf{S}; \mathbf{D}) = P(\mathbf{S}|\mathbf{D})P_0(\mathbf{D})$. Here $P_0(\mathbf{D})$ is the prior probability distribution of diseases, which could depend on the patient's characteristics such as gender and age and disease properties such as duration of a disease, mortality rate and transmission rate among others.

Let $P_{true}(\mathbf{S}|\mathbf{D})$ be the true probability distribution of sign variables given disease hypothesis **D**. In practice, we may have access only to a small subset of marginal probabilities of this true distribution. For instance, suppose we are given sign probabilities $P_{true}(S_i|\text{nodisease})$, $P_{true}(S_i, S_j|\text{only}D_a)$, and $P_{true}(S_i, S_j|\text{only}D_a, D_b)$ conditioned on the absence of any of the diseases, and the presence of only one and two diseases, respectively. Using the maximum entropy principle[30], for the conditional probability distribution of signs we take[23]

$$P(\mathbf{S}|\mathbf{D}) = \frac{1}{Z(\mathbf{D})}\phi_0(\mathbf{S}) \times \prod_a \phi_a(\mathbf{S}|D_a) \times \prod_{a<b} \phi_{ab}(\mathbf{S}|D_a, D_b), \tag{1}$$

where the partition function $Z(\mathbf{D})$ is obtained from normalization $\sum_{\mathbf{S}} P(\mathbf{S}|\mathbf{D}) = 1$. More precisely, the disease interaction factors $(\phi_0, \phi_a, \phi_{ab})$, are given by

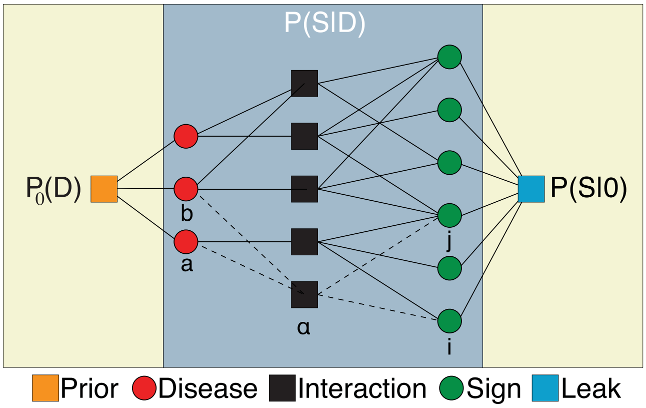

Figure 2. The interaction graph of disease variables (left circles) and sign variables (right circles) related by M_a one-disease and M_{ab} two-disease interaction factors (middle squares) in addition to interactions induced by the leak probability (right square) and the prior probability of diseases (left square). In general, an interaction factor $\alpha = a, ab$ is connected to k_α signs and l_α diseases[23].

$$\phi_0(\mathbf{S}) \equiv e^{\sum_i K_i^0 S_i}, \tag{2}$$

$$\phi_a(\mathbf{S}|D_a) \equiv e^{D_a\left[\sum_i K_i^a S_i + \sum_{i<j} K_{ij}^a S_i S_j\right]}, \tag{3}$$

$$\phi_{ab}(\mathbf{S}|D_a, D_b) \equiv e^{D_a D_b\left[\sum_i K_i^{ab} S_i + \sum_{i<j} K_{ij}^{ab} S_i S_j\right]}. \tag{4}$$

In principle, the above information from the true probability distribution is sufficient to determine the model parameters K_i^0, $K_i^{a,ab}$, and $K_{ij}^{a,ab}$. Figure 2 shows the interaction graph of sign and disease variables related by the above interaction factors[23]. We use M_a and M_{ab} for the number of one-disease and two-disease interaction factors, respectively. An interaction factor is connected to k_a or k_{ab} signs depending on the number of involved diseases.

Simplifying Assumptions. For simplicity, in the main text, we ignore the sign-sign interactions in the interaction factors ($K_{ij}^a = K_{ij}^{ab} = 0$). That is, we consider the one-disease-one-sign (D1S1) model with parameters K_i^0, K_i^a, and two-disease-one-sign (D2S1) model with parameters K_i^0, K_i^a, K_i^{ab}. This allows us to compute exactly the partition function for these models. Moreover, given the true marginals, the parameters K_i^0, $K_i^{a,ab}$ of the D1S1 and D2S1 models can be computed exactly. To be specific, in the main text we focus on the D2S1 model, which works well as long as the number of present diseases in the hypothesis, $|\mathbf{D}|$, is less than or equal to two[23]. We shall briefly discuss the results obtained from the simpler D1S1 model and the more difficult D2S2 model (including the two-sign interactions) in Supplementary Information, Appendixes A and B, respectively.

In addition, we assume that the prior disease probability is factorized, $P_0(\mathbf{D}) = \prod_{a=1}^{N_D} P_0(D_a)$, with $P_0(D_a) = e^{K_a^0 D_a}/(1 + e^{K_a^0})$. The parameters K_a^0 can be used to control the expected number of present diseases in the hypothesis. For instance, K_a^0 can be chosen such that $N_D P_0(D_a = 1) = |\mathbf{D}|$. Alternatively, we can fix the expected number of disease probabilities which are greater than a threshold value. As long as the number of signs and diseases is small (e.g., $N_S = 20$, $N_D = 5$), we work with a fully connected model of the variables, where all the one-disease and two-disease interaction factors (ϕ_a, ϕ_{ab}) including the interactions with all the sign variables could be present in the model. The graph parameters defining the structure of a fully connected model are: $M_a = N_D$, $M_{ab} = 0$, $k_a = N_S$ in the D1S1 model and $M_a = N_D$, $M_{ab} = N_D(N_D - 1)/2$, $k_a = k_{ab} = N_S$ in the D2S1 model. For larger number of variables, we limit ourselves to sparsely connected graphs with smaller number of interaction factors (M_a, M_{ab}) and connectivities (k_a, k_{ab}).

Diagnosis. Let us assume that a subset $\mathbf{I}_0 = \{i_1, i_2, \ldots, i_{N_O}\}$ of the sign variables is observed with values $\mathbf{S}^o$. The possible values for the remaining subset of unobserved signs are denoted by $\mathbf{S}^u$. At each time step $t = 1, 2, \ldots, T$ we use a strategy to choose one of the unobserved signs j_t for observation. The sequence of observed signs at time step t is represented by $\mathbf{O}(t) = \mathbf{I}_0 \cup \{j_1, \ldots, j_t\}$. We use $\mathbf{U}(t)$ for the subset of unobserved signs.

At each step we have the disease and sign probabilities,

$$P(\mathbf{D}|\mathbf{S}^o) \propto \sum_{\mathbf{S}^u} P(\mathbf{S}|\mathbf{D}) P_0(\mathbf{D}), \tag{5}$$

$$P(\mathbf{S}^u|\mathbf{S}^o) \propto \sum_{\mathbf{D}} P(\mathbf{S}|\mathbf{D}) P_0(\mathbf{D}), \tag{6}$$

which can be used to compute the disease marginal probabilities $P(D_a|\mathbf{S}^o)$ and the sign marginal probabilities $P(S_i|\mathbf{S}^o)$. The maximum likelihood (ML) hypothesis $\mathbf{D}^{ML}$ is obtained by maximizing the disease likelihood[29],

$$\mathcal{L}(\mathbf{D}|\mathbf{S}^o) \equiv \sum_{\mathbf{S}^u} P(\mathbf{S}|\mathbf{D})P_0(\mathbf{D}). \tag{7}$$

At each step t we choose an unobserved sign for observation which maximizes an appropriate objective function of the chosen sign. A reasonable objective function is the maximum value of the disease likelihood,

$$\mathrm{ML}(t) \equiv \frac{1}{|\mathbf{O}(t)|}\langle \log \mathcal{L}(\mathbf{D}^{ML}|\mathbf{S}^o(t))\rangle_O. \tag{8}$$

The average $\langle\cdot\rangle_O$ in the above equations is taken over the probability distribution of observation outcomes. Note that before the medical observation we only know the marginal probability of the chosen sign $P(S_j|\mathbf{S}^o(t-1))$. And, after each observation (medical test), we obtain the true value of the observed sign.

We assume that the aim of the diagnostic process is to reach the correct diagnosis with the minimum number of medical tests. Obviously, a disease probability that is closer to zero or one could be more helpful to decide if the disease is absent or present. Therefore, we may at each step choose the sign that results to the largest polarization of the disease probabilities:

$$\mathrm{DP}(t) \equiv \frac{1}{\sum_a W_a}\sum_a W_a \left\langle \left| P(D_a{=}1|\mathbf{S}^o(t)) - \frac{1}{2} \right| \right\rangle_O. \tag{9}$$

Other measures of polarization, e.g., the root-mean-square of single-disease polarizations, may work as well[23]. Here we are taking into account also the importance or weight of the diseases W_a, which could be high for example for life threatening diseases. The $P(D_a=1|\mathbf{S}^o(t))$ give the disease probabilities after the t-th observation. The marginal probabilities are obtained from the reconstructed models of the true probability distribution.

In this paper, however, we are interested in simulation of the above sequential process of decisions and observation for T steps, without asking for any real medical test to reveal the true sign values. In other words, we are interested in extrapolation or prediction of the diagnostic process starting from a small subset $\mathbf{I}_0$ of the observed signs and a simple model of the sign and disease variables. Here, an observed sign j in the process is treated as a stochastic variable with a value that is sampled from the associated marginal probability $P(S_j|\mathbf{S}^o(t-1))$ at that time step. For brevity, we call this type of diagnosis Diags-II, and Diags-I is used to refer the diagnostic process in which the true sign value is known (by medical test) just after choosing the sign for observation.

More precisely, in the case of Diags-I, at each time step t we choose an unobserved sign j_t, which maximizes the following objective function

$$\mathcal{E}(t) \equiv \mathrm{ML}(t) + \lambda_P \mathrm{DP}(t) - \lambda_C SC(t). \tag{10}$$

Then we do the medical test to find out the true value of the chosen sign, and go to the next step of the diagnostic process. We have included also the sign cost $SC(t) \equiv C_{j_t}$ into the objective function. The λ_P and λ_C are parameters to control the degree of disease polarization and cost of the observations, respectively. In the case of Diags-II, we choose an optimal sequence of decisions $\mathbf{O}(T)$, which maximizes the following objective functional of the candidate observations:

$$\mathcal{E}[\mathbf{O}(T)] \equiv \sum_{t=1}^{T}\mathrm{ML}(t) + \lambda_P\sum_{t=1}^{T}\mathrm{DP}(t) - \lambda_C\sum_{t=1}^{T} SC(t). \tag{11}$$

Simplifying Assumptions. A greedy approximation of Diags-II is obtained by splitting the whole process into T independent steps; this is very similar to Diags-I except the fact that here we do not know the true sign values. But, we have an estimate of the marginal sign probabilities $P(S_j|\mathbf{S}^o)$, which can be utilized to assign a good value to the "observed" sign. The time complexity of the optimization algorithm is then of order $(N_S - N_O)T$ times the complexity of computing the marginal sign/disease probabilities and the maximum likelihood. These computations can be done by approximate inference and optimization algorithms based on the Monte Carlo sampling. For sparse interaction graphs of sign and disease variables, the time complexity of such an algorithm would be proportional to N_S.

To simplify the study and reduce the computation time, we shall replace the average over the possible realizations of the observation outcome with the most probable value. Suppose that we are to observe sign j at time step t. Then, we assume that the outcome of each observation is given by the value which maximizes the corresponding marginal probability at that time step, i.e., $S_j = \arg\max P(S_j|\mathbf{S}^o)$.

Diagnostic Performance Measures.

The main question of this study is: How close are the predictions obtained by Diags-II to the (more expensive) Diags-I? And, when we can trust the outcome of such a diagnostic process? More precisely, given the model of sign and disease variables, we shall see how predictions of the Diags-II improve by increasing the number of initial observations. This of course depends on the quality of the reconstructed models, the structure of the objective function and performance of the optimization algorithm which is used in the study of Diags-II, and the number of initial observed signs.

To check the quality of our extrapolation, we shall take a simple benchmark model for the true probability distribution $P_{true}(\mathbf{S}|\mathbf{D})$. Given, any disease hypothesis $\mathbf{D}^{true}$, the associated signs $\mathbf{S}^{true}$ can then be obtained by taking the most probable signs from the true probability distribution. To be specific, for the true model we take the following exponential distribution:

$$P_{true}(\mathbf{S}|\mathbf{D}) = \frac{1}{Z_{true}(\mathbf{D})}e^{-H(\mathbf{S},\mathbf{S}(\mathbf{D}))}, \tag{12}$$

where the Hamming distance $H(\mathbf{S}, \mathbf{S}') = \sum_i (S_i - S'_i)^2/4$ gives the number of different signs in the two sign configurations. Here $\mathbf{S}(\mathbf{D})$ defines the signs attributed to $\mathbf{D}$. We will choose these signs randomly and uniformly from the configuration space of sign variables.

Consider the diagnostic process for a patient with true disease values $\mathbf{D}^{true}$. At any time step t, we compute the overlap of the disease probabilities with the true disease hypothesis,

$$DL(t) \equiv \frac{1}{\sum_a W_a}\sum_a W_a(2D_a^{true} - 1)\left(P(D_a = 1|\mathbf{S}^o(t)) - \frac{1}{2}\right). \tag{13}$$

This shows how well the inferred disease probabilities are close to the true disease values. Obviously, $DL(t)$ always increases (on average) with the number of observations in the Diags-I. But this quantity can decrease or increase with t depending on number of initial observations $N_O(0)$.

Another interesting quantity is the first diagnosis time for a specific subset of diseases $\mathbf{A}$; we define the first right diagnosis time T_R as the first time at which we find:

$P(D_a = 1|\mathbf{S}^o(t)) \geq P_{th}$ for at least one of the diseases $a \in \mathbf{A}$.

In the same way we define the first wrong diagnosis time T_W as the first time at which:

$P(D_a = 1|\mathbf{S}^o(t)) \geq P_{th}$ for at least one of the diseases $a \notin \mathbf{A}$.

Then, the probability of having right or wrong diagnosis after t observations would critically depend for example on the initial number of observations.

Simplifying Assumptions. To obtain an upper bound for the critical number of initial observations, we use a random strategy for suggesting the observations in the diagnostic process. By the random strategy we mean that at each step we choose randomly an unobserved sign for the next observation. Then, in the Diags-I (random), we do a real observation to find out the true value of the chosen sign. Instead, in the Diags-II (random), we assign the most probable value of the sign to the suggested sign for observation and go ahead without doing any real observation.

Approximation Algorithms

A (zero-temperature) Monte Carlo algorithm. In the following, we shall work with the sequence configuration $\mathbf{I}_{1\rightarrow T} \equiv \{j_1, \ldots, j_T\}$ instead of the whole set of observations $\mathbf{O}(T) = \mathbf{I}_0 \cup \{j_1, \ldots, j_T\}$. For any such configuration, we can compute the marginal probabilities $P(S_j|\mathbf{S}^o(t))$ and $P(D_a|\mathbf{S}^o(t))$, and the objective function $\mathcal{E}[\mathbf{I}_{1\rightarrow T}]$, by an exact algorithm (for small number of variables) or an approximate algorithm (for larger number of variables). In either case, we have to run the algorithm for T times to compute the disease probabilities conditioned on the values of the observed signs in the previous steps. Thus, the time complexity of the algorithm is proportional to T times the time complexity of computing the objective function[31]. The main steps of the optimization algorithm are:

- Input: the model $P(\mathbf{S}; \mathbf{D})$, the weights W_a and costs C_i, the parameters $\lambda_{P,C}$, initial set of observed signs $\mathbf{I}_0$, time steps T
- Start from an initial (random) sequence of observations $\mathbf{I}_{1\rightarrow T} = \{j_1, \ldots, j_T\}$:
- compute the objective function $\mathcal{E}[\mathbf{I}_{1\rightarrow T}]$
- For $n = 1, 2, \ldots, n_{max}$:
  - suggest a new configuration $\mathbf{I}'_{1\rightarrow T}$
  - compute the change $\Delta\mathcal{E} = \mathcal{E}[\mathbf{I}'_{1\rightarrow T}] - \mathcal{E}[\mathbf{I}_{1\rightarrow T}]$ in the objective function
  - accept the new configuration if $\Delta\mathcal{E} > 0$
- Output: the (local) optimal configuration $\mathbf{I}^{opt}_{1\rightarrow T}$

Computing the objective function and generating a new sequence configuration $\mathbf{I}'_{1\rightarrow T}$ from $\mathbf{I}_{1\rightarrow T}$ are the main parts of the algorithm. In a previous study[23], we found that a good heuristic strategy is to choose at each step the most positive unobserved sign for the next observation. The most positive sign is the one with the maximum probability of being positive, that is $i_{MP} = \arg\max_{i\in U} P(S_i = +1|\mathbf{S}^o)$. It is important that the assigned values are as close as possible to the true values. By choosing the most positive signs we indeed try to reduce the error in prediction of the values of the observed signs in the simulation process. Note that a wrong assignment at the early stages of the process can significantly affect the whole process, consequently affecting the diagnosis.

Here we use this finding to guide the updating step of the optimization algorithm. More specifically, we use the following rules to update a sequence configuration $\mathbf{I}_{1\rightarrow T}$:

- choose randomly a time step $1 \leq \tau \leq T$
- for $t = \tau, \ldots T$, suggest an unobserved sign $j_{t'}$ with a probability proportional to $P(S_{j_{t'}} = +1|\mathbf{S}^o(t - 1))$

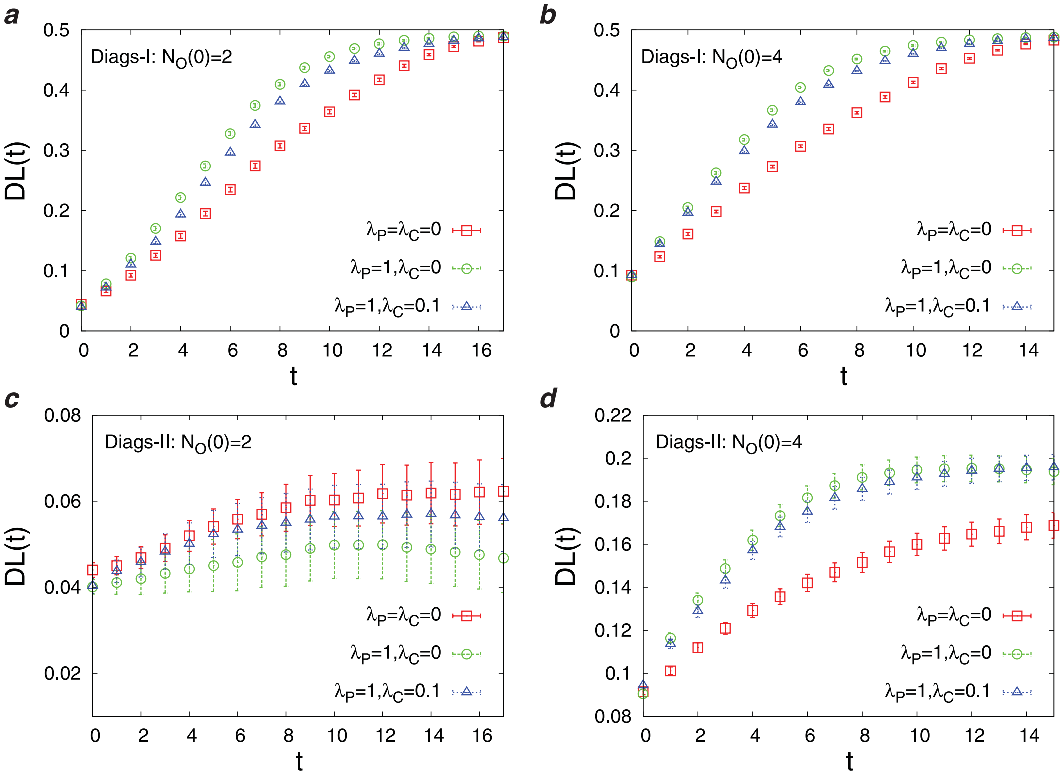

Figure 3. Dependence of $DL(t)$ on the initial number of observed signs $N_O(0)$ and the parameters λ_P, λ_C. The results have been obtained from the D2S1 model by (**a**)–(**b**) the Diags-I (greedy), and (**c**)–(**d**) the Diags-II (greedy) with the prior probabilities $P_0(D_a=1)=2/N_D$. The model parameters of the (fully connected) D2S1 model are obtained exactly from the conditional marginals of the true exponential model. A disease hypothesis is chosen randomly for the simulation with a probability proportional to the weights of the present diseases. All the marginal probabilities have been computed exactly for a small number of sign and disease variables ($N_S=20$ and $N_D=5$). The data are results of averaging over at least 500 independent realizations of the models and simulation process.

The above process suggests the new sequence $\mathbf{I}'_{1\to T}=\{j_1, \ldots, j_{\tau-1}, j_\tau', \ldots, j'_T\}$ which is accepted only if the new sequence increases the objective function. The success probability of a candidate sequence suggested in this way is about 0.57 (in 660 trials) for the Diags-II, with $N_S=500$, $N_D=50$, $N_O(0)=50$, and $T=50$. Details of computing the objective function is given in Methods Section. Very briefly, to compute the objective function we need the sign and disease marginal probabilities (for $DP(t)$), which are estimated by a standard Monte Carlo algorithm, and the maximum likelihood value (for $ML(t)$), which is estimated by a Simulated Annealing algorithm[25]. The latter computation can again be done by a zero-temperature Monte Carlo algorithm, but since it determines the objective function we prefer to employ a more accurate optimization algorithm. The time complexity of these algorithms in a sparse D2S1 model is proportional to the number of diseases.

Results

In this section, we present the results obtained by the numerical simulations of the Diags-I and Diags-II for different parameters in the objective function (λ_P, λ_C) and different number of initial observations $N_O(0)$. In Fig. 3 we report the overlap of the disease probabilities with the true disease values, $DL(t)$, as the number of observations t increases starting from an initial number of observations. Here, we observe the impact of disease polarization and initial observations on $DL(t)$ using the greedy strategy. Figure 4 displays the joint probability distribution of the first diagnosis times $P(T_R,T_W)$ in the D2S1 model with the Diags-II (greedy). To see better the effects of (λ_P, λ_C) and $N_O(0)$ on the first diagnosis times, in Figs 5 and 6 we show the cumulative probability distributions $P(T_R\leq t)$ and $P(T_W\leq t)$. It is important to know how much the parameter λ_C reduces the cost of observations $SC(t)$. Figure 7 displays the cumulative cost $\sum_{t'=1}^{t}SC(t')$ for two values of λ_C we used in the numerical simulations. The number of variables in these figures is sufficiently small ($N_S=20$, $N_D=5$), which allows us to compute exactly the marginal sign/disease probabilities by an exhaustive sampling algorithm. To check the results for larger problem sizes, we have to resort to the Monte Carlo algorithms introduced in the previous sections. Figures 8–10 show the behavior of the first diagnosis times for the D2S1 model using the Diags-II. Here, we compare the results obtained by a random strategy with those that are obtained by maximizing the objective function.

Main points of this study are:

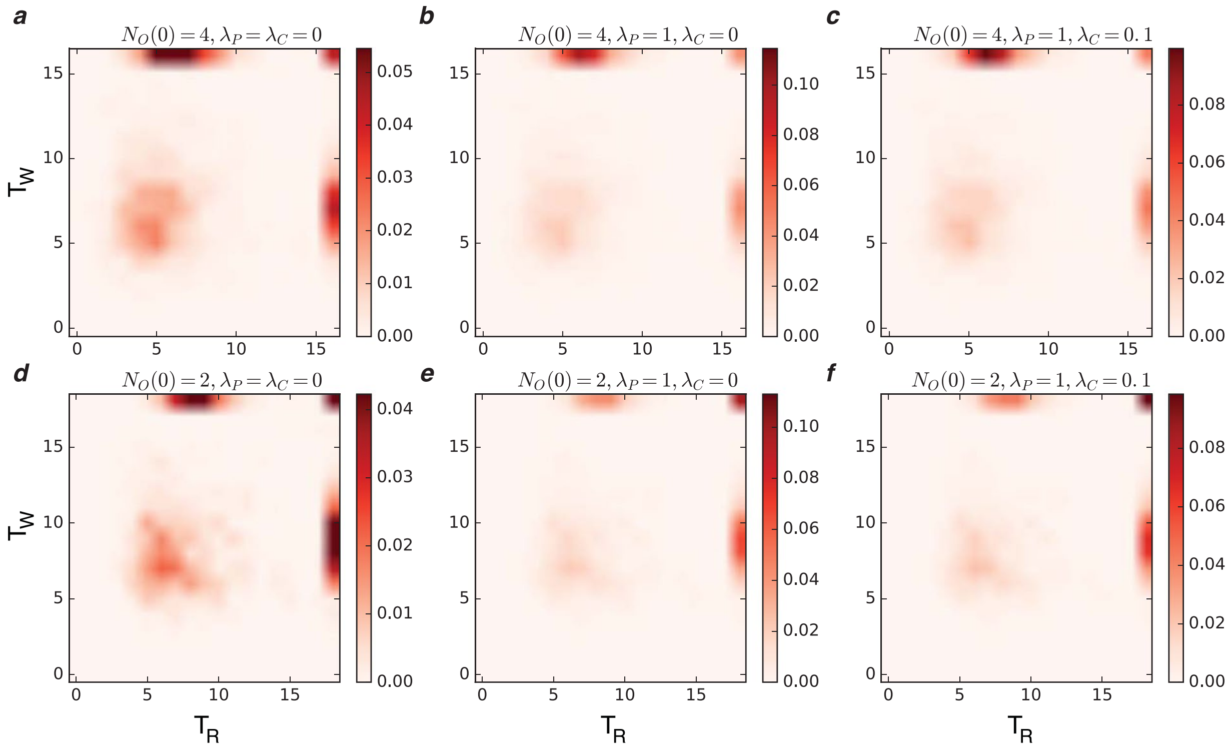

Figure 4. The joint probability of the first diagnosis times, $P(T_R, T_W)$, for different numbers of initial observations $N_O(0)$ and the parameters λ_P, λ_C. The panels display the cases: (**a**)–(**c**) $N_O(0) = 4$ and (**d–f**) $N_O(0) = 2$ for ($\lambda_P = \lambda_C = 0$), ($\lambda_P = 1$, $\lambda_C = 0$), ($\lambda_P = 1$, $\lambda_C = 0.1$), respectively. The results have been obtained from the D2S1 model by the Diags-II (greedy) with the prior probabilities $P_0(D_a = 1) = 2/N_D$, and the threshold probability $P_{th} = 0.9$. The last values of T_R and T_W are reserved for the case in which the corresponding disease probabilities remain less than the threshold value during the whole process. The model parameters of the (fully connected) D2S1 model are obtained exactly from the conditional marginals of the true exponential model. A disease hypothesis is chosen randomly for the simulation with a probability proportional to the weights of the present diseases. The number of present diseases in the hypothesis is $|\mathbf{D}| = 2$. All the marginal probabilities have been computed exactly for a small number of sign and disease variables ($N_S = 20$ and $N_D = 5$). The data are results of at least 500 independent realizations of the model and simulation process.

- Figs 3–6 show that adding a measure of disease polarization to the standard objective function (the maximum value of the log-likelihood) improves the diagnostic performance.
- We find that the cost of observations can considerably be decreased without seriously affecting the diagnostic performance. As Figs 3–7 show, the overlap of the disease probabilities with the true disease values does not significantly change by considering a small penalty for the cost of observations in the objective function. This is the case specially for intermediate values of t, where the optimized cumulative cost displays the largest deviation from the unoptimized one.
- The diagnostic performance of the Diags-I process always increases with the number of observations, because any observation (even if suggested randomly) reveals the true value of a previously unobserved sign. However, we see in Figs 3 and 6 that the performance of the Diags-II depends critically on the initial number of the observed signs. In particular, for $N_O(0) < N_O^*$ the Diags-II process more likely results in a wrong diagnosis, with $P(T_R \leq T) - P(T_W \leq T) < 0$ for a sufficiently large number of observations T. On the other hand, the probability of a right diagnosis in the Diags-II process is greater than the wrong one for $N_O(0) > N_O^*$. We obtain an upper bound for this critical value of N_O^* for a sparse model of sign and disease variables, with $N_D = 50$, $N_S = 500$, $M_a = 50$, $M_{ab} = 100$, $k_a = k_{ab} = 150$, see Figs 8 and 9. The upper bound is obtained by the random strategy where at each step we choose randomly and uniformly an unobserved sign for observation. In general, we expect that such a critical value to be proportional to the total number of signs in the model, and of course dependent on the model structure. Here, the marginal sign/disease probabilities are computed by a standard Monte Carlo algorithm.
- Moreover, even for $N_O(0) < N_O^*$, there exists a characteristic number of observations t*, where for t < t* the probability of inferring the right diseases in the Diags-II is still larger than that of the wrong diagnosis (Fig. 9). The characteristic time t* of course increases with the number of initial observations. In other words, t* gives the maximum number of observation tests (in the Diags-II) we can choose randomly before missing the useful information provided by the initial observations.
- We use the marginal sign probabilities $P(S_j = +1|\mathbf{S}^o)$ to guide the updating step of a (zero-temperature) Monte Carlo algorithm for optimizing the objective functional of the observations. This algorithm was used

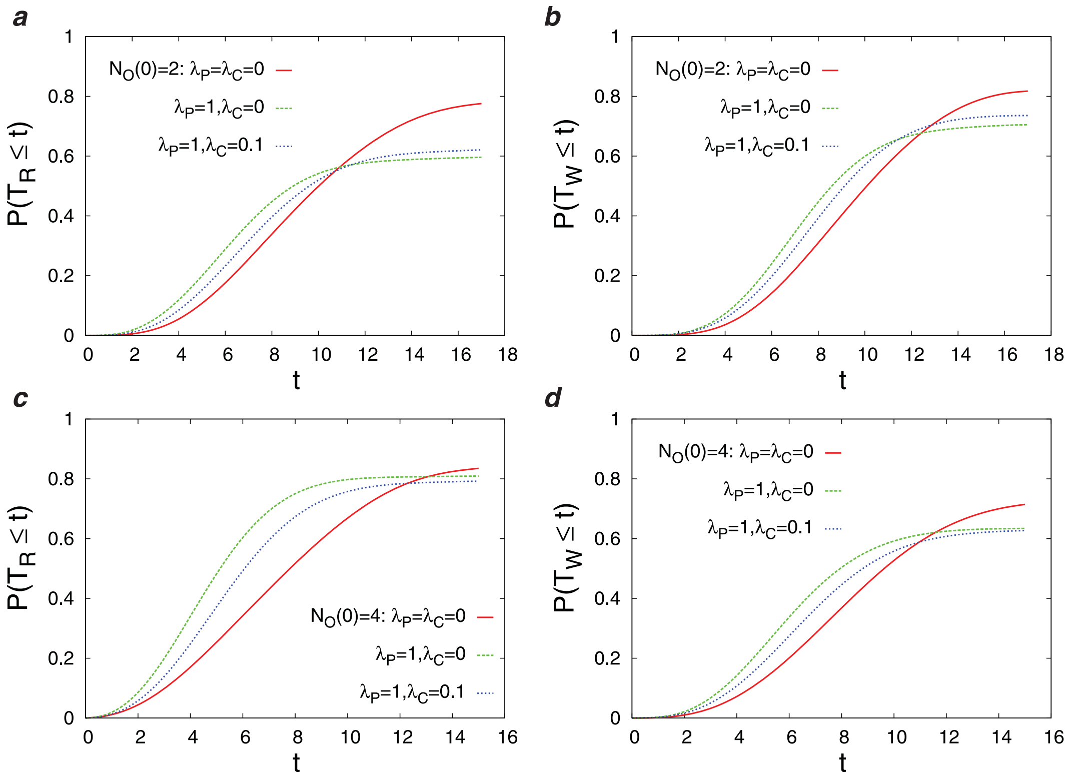

Figure 5. The cumulative probabilities $P(T_R \le t)$ and $P(T_W \le t)$ of the first diagnosis times for different numbers of the initial observations $N_O(0)$ and the parameters λ_P, λ_C. The results have been obtained from the D2S1 model by the Diags-II (greedy) with the prior probabilities $P_0(D_a = 1) = 2/N_D$, and the threshold probability $P_{th} = 0.9$. The model parameters of the (fully connected) D2S1 model are obtained exactly from the conditional marginals of the true exponential model. A disease hypothesis is chosen randomly for the simulation with a probability proportional to the weights of the present diseases. The number of present diseases in the hypothesis is $|\mathbf{D}| = 2$. All the marginal probabilities have been computed exactly for a small number of sign and disease variables ($N_S = 20$ and $N_D = 5$). The data are results of at least 500 independent realizations of the model and simulation process.

to uncover a small number (one or two) of hidden diseases in sparsely interacting models of signs and diseases, by simulating the Digas-II process. Figure 10 compares the first diagnosis times (T_R, T_W) obtained by the above algorithm with the ones predicted by the random strategy. Here, the marginal sign/disease probabilities and the maximum of the log-likelihood in the objective function are computed by the standard Monte Carlo and Simulated Annealing algorithms.

Discussions and Conclusions

In this article, using novel, simple and rigorous models, we studied the efficiency determinants of a diagnostic approach, i.e. choice of a sequence of steps that leads to a diagnosis. We assessed the tradeoff between the number of steps (tests) and the cost (financial cost and biological risk) involved. We compared the efficiency of a sequential step-by-step diagnostic approach (i.e. a medical test is ordered and then the next test is decided) with an approach that orders a batch of tests at once during a clinical session. We recommend a combination of the two approaches, i.e. starting with the step-by-step approach and then switching to the batch approach would be optimal. The timing of when to switch is then dependent on the collected mass of information. At a certain critical point, switching the strategy would allow for faster clinical management. Moreover, we defined and reflected on an inherent property of a test, termed as disease polarization, that needs to be considered in constructing an efficient diagnostic flowchart. Our model includes interactions between diseases (and signs) which are typically neglected in the literature, but are emerging as important ingredients in omics analyses of human physiology and diseases[32].

Finally, it should be mentioned that the typical diagnostic problems may involve many differentials (e.g. a few hundreds or thousands of diseases and signs)[13]. Monte Carlo is a computationally extensive algorithm to deal with large-scale problems. However, it works well independent of the model structure, if provided with adequate time. In our previous work, we proposed an approximate algorithm that is based on the Bethe approximation, but it works well for very sparse interaction graphs[23]. In a recent work, we are going to use the mean-field approximation, which again works well in fully-connected interaction graphs (unpublished data). Of course, the algorithms that are based on Bethe and mean-field approximations are more efficient than Monte Carlo. But, as mentioned earlier, their performance is limited by the structure of the model.

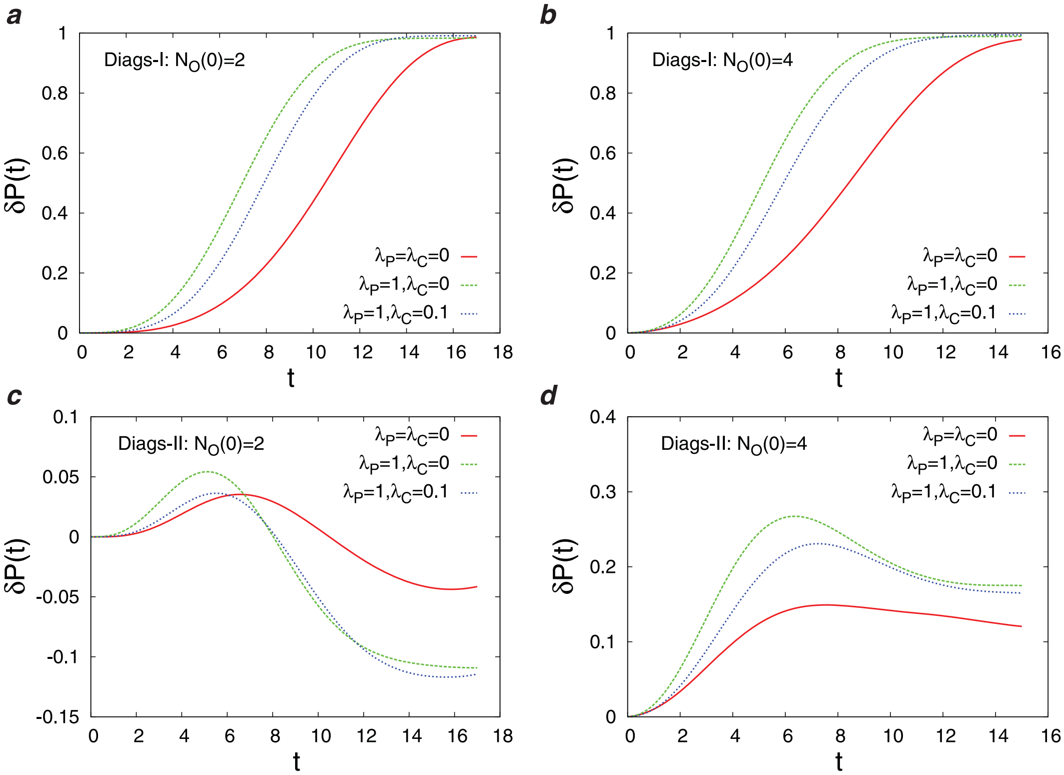

Figure 6. The difference $\delta P(t) \equiv P(T_R \le t) - P(T_W \le t)$ in the cumulative probabilities of the first diagnosis times for different numbers of the initial observations $N_O(0)$ and the parameters λ_P, λ_C. The results have been obtained from the D2S1 model by (**a**),(**b**) the Diags-I (greedy), and (**c**)–(**d**) Diags-II (greedy) with the prior probabilities $P_0(D_a = 1) = 2/N_D$, and the threshold probability $P_{th} = 0.9$. The model parameters of the (fully connected) D2S1 model are obtained exactly from the conditional marginals of the true exponential model. A disease hypothesis is chosen randomly for the simulation with a probability proportional to the weights of the present diseases. The number of present diseases in the hypothesis is $|\mathbf{D}| = 2$. All the marginal probabilities have been computed exactly for a small number of sign and disease variables ($N_S = 20$ and $N_D = 5$). The data are results of at least 500 independent realizations of the model and simulation process.

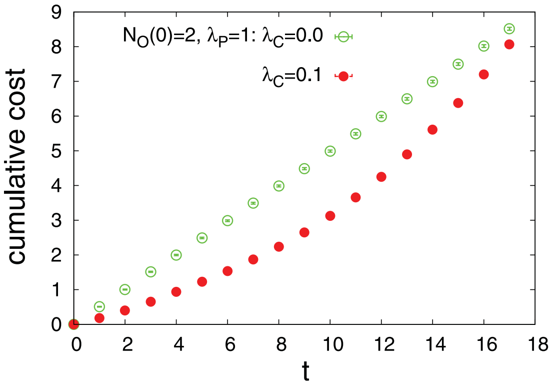

Figure 7. Cumulative cost of the diagnosis for two different values of λ_C. The results have been obtained from the D2S1 model by the Diags-II (greedy) with the prior probabilities $P_0(D_a = 1) = 2/N_D$. The model parameters of the (fully connected) D2S1 model are obtained exactly from the conditional marginals of the true exponential model. A disease hypothesis is chosen randomly for the simulation with a probability proportional to the weights of the present diseases. All the marginal probabilities have been computed exactly for a small number of sign and disease variables ($N_S = 20$ and $N_D = 5$). The data are results of averaging over at least 500 independent realizations of the model and simulation process.

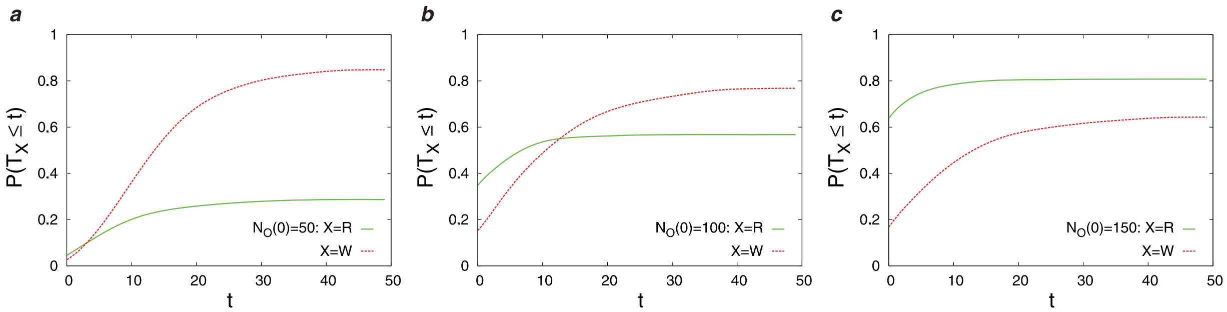

Figure 8. The cumulative probabilities $P(T_R \leq t)$ and $P(T_W \leq t)$ of the first diagnosis times for different numbers of the initial observations $N_O(0)$. The results have been obtained from a sparse D2S1 model by the Diags-II (random) with the prior probabilities $P_0(D_a = 1) = 2/N_D$, and the threshold probability $P_{th} = 0.9$. The model parameters of the (sparse) D2S1 model are obtained exactly from the conditional marginals of the true exponential model. The interaction graph and model parameters are: $N_S = 500$, $N_D = 50$, $M_a = 50$, $M_{ab} = 100$, $k_a = 150$, $k_{ab} = 150$. A disease hypothesis is chosen randomly for the simulation with a probability proportional to the weights of the present diseases. The number of present diseases in the hypothesis is $|\mathbf{D}| = 2$. The marginal probabilities have been computed approximately by the Monte Carlo algorithm. The data are results of at least 200 independent realizations of the model and simulation process.

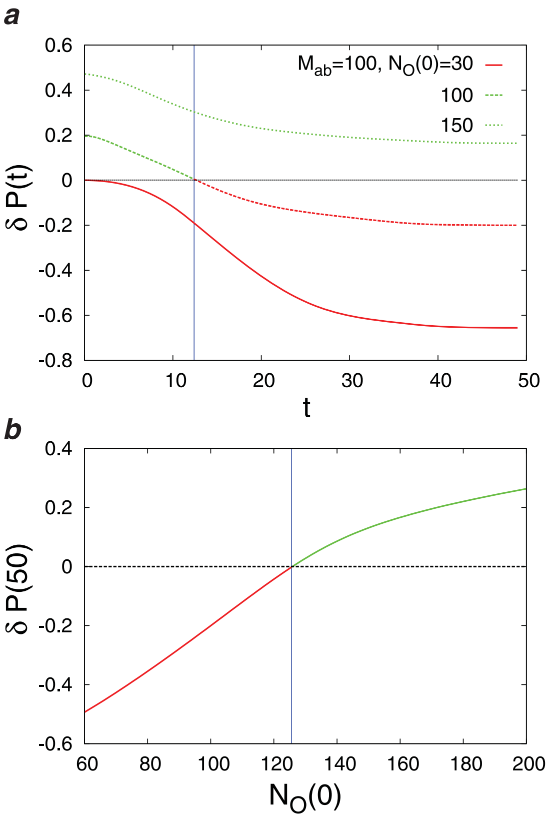

Figure 9. The difference $\delta P(t) \equiv P(T_R \leq t) - P(T_W \leq t)$ in the cumulative probabilities of the first diagnosis times: (**a**) vs the number of observations t for different numbers of the initial observations $N_O(0)$, and (**b**) $\delta P(50)$ vs $N_O(0)$ for a sufficiently large value of t. The results have been obtained from a sparse D2S1 model by the Diags-II (random) with the prior probabilities $P_0(D_a = 1) = 2/N_D$, and the threshold probability $P_{th} = 0.9$. The model parameters of the (sparse) D2S1 model are obtained exactly from the conditional marginals of the true exponential model. The interaction graph and the model parameters are: $N_S = 500$, $N_D = 50$, $M_a = 50$, $M_{ab} = 100$, $k_a = 150$, $k_{ab} = 150$. A disease hypothesis is chosen randomly for the simulation with a probability proportional to the weights of the present diseases. The number of present diseases in the hypothesis is $|\mathbf{D}| = 2$. The marginal probabilities have been computed approximately by the Monte Carlo algorithm. The data are results of at least 200 independent realizations of the model and simulation process.

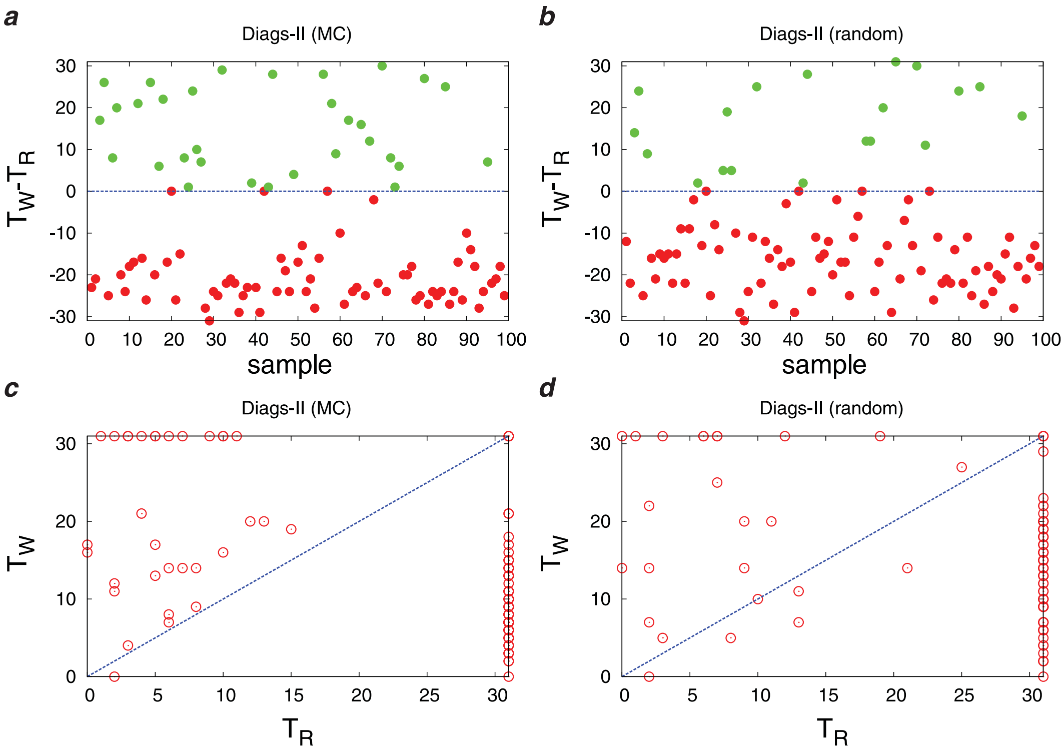

Figure 10. (**a**),(**b**) The difference $T_W - T_R$ in the first diagnosis times, and (**c**),(**d**) T_W vs T_R for some independent realizations of the problem. The results have been obtained from a sparse D2S1 model by the Diags-II (zero-temperature MC) in panels (**a**)–(**c**), and the Diags-II (random) in panels (**b**)–(**d**). The prior probabilities are $P_0(D_a = 1) = 2/N_D$, and the threshold probability is $P_{th} = 0.9$. The model parameters of the (sparse) D2S1 model are obtained exactly from the conditional marginals of the true exponential model. The interaction graph and the model parameters are: $N_S = 500$, $N_D = 50$, $M_a = 50$, $M_{ab} = 100$, $k_a = 150$, $k_{ab} = 150$. The algorithms are given $N_O(0) = 50$ initial observations to suggest a sequence of $T = 30$ other observations for diagnosis. Here we take $\lambda_P = 1$ and $\lambda_C = 0$. A disease hypothesis is chosen randomly for the simulation with a probability proportional to the weights of the present diseases. The number of present diseases in the hypothesis is $|\mathbf{D}| = 2$. The marginal probabilities have been computed approximately by the Monte Carlo algorithm.

Method

Computing the objective function.

Here we consider only the $D1S1$ and $D2S1$ models, where we can exactly compute the partition function

$$Z(\mathbf{D}) = \prod_i \left(2\cosh\left[K_i^0 + \sum_a K_i^a D_a + \sum_{a<b} K_i^{ab} D_a D_b\right]\right). \tag{14}$$

For these models, we can also exactly compute the model parameters given the probabilities $P_{true}(S_i|\text{nodisease})$, $P_{true}(S_i|\text{only } D_a)$, and $P_{true}(S_i|\text{only } D_a, D_b)$,

$$K_i^0 = \frac{1}{2}\ln\left(\frac{P_{true}(S_i = +1|\text{nodisease})}{P_{true}(S_i = -1|\text{nodisease})}\right), \tag{15}$$

$$K_i^a = \frac{1}{2}\ln\left(\frac{P_{true}(S_i = +1|\text{only } D_a)}{P_{true}(S_i = -1|\text{only } D_a)}\right) - K_i^0, \tag{16}$$

$$K_i^{ab} = \frac{1}{2}\ln\left(\frac{P_{true}(S_i = +1|\text{only } D_a, D_b)}{P_{true}(S_i = -1|\text{only } D_a, D_b)}\right) - K_i^0 - K_i^a - K_i^b. \tag{17}$$

For a given subset $\mathbf{O}$ of observed signs with values $\mathbf{S}^o$, the disease probabilities are obtained from

$$P(D_a = 1|\mathbf{S}^o) = \frac{1}{\mathcal{Z}(\mathbf{D}|\mathbf{S}^o)}\sum_{\mathbf{D}} D_a e^{-\mathcal{H}(\mathbf{D}|\mathbf{S}^o)}, \tag{18}$$

where $-\mathcal{H}(\mathbf{D}|\mathbf{S}^o) \equiv \log\mathcal{L}(\mathbf{D}|\mathbf{S}^o)$ is the log-likelihood function

$$\mathcal{H}(\mathbf{D}|\mathbf{S}^o) = -\sum_a K_a^0 D_a - \sum_{i\in\mathbf{O}} S_i^o h_i(\mathbf{D}) + \sum_{i\in\mathbf{O}} \ln(2\cosh h_i(\mathbf{D})), \tag{19}$$

and $\mathcal{Z}(\mathbf{D}|\mathbf{S}^o)$ is a normalization constant,

$$\mathcal{Z}(\mathbf{D}|\mathbf{S}^o) \equiv \sum_{\mathbf{D}} e^{-\mathcal{H}(\mathbf{D}|\mathbf{S}^o)}. \tag{20}$$

For brevity, here we defined the local field experienced by sign i as

$$h_i(\mathbf{D}) \equiv K_i^0 + \sum_a K_i^a D_a + \sum_{a<b} K_i^{ab} D_a D_b. \tag{21}$$

It is easy to show that the marginal probability of an unobserved sign is given by:

$$P(S_i = 1|\mathbf{S}^o) = \frac{1}{\mathcal{Z}(\mathbf{D}|\mathbf{S}^o)} \sum_{\mathbf{D}} \left(\frac{1 + \tanh h_i(\mathbf{D})}{2}\right) e^{-\mathcal{H}(\mathbf{D}|\mathbf{S}^o)}. \tag{22}$$

Now, we can use the standard Monte Carlo algorithm with the energy function $\mathcal{H}(\mathbf{D}|\mathbf{S}^o)$, to compute the marginal probabilities which are needed for the objective function. More precisely, we need to sample the disease configurations with a probability proportional to $\exp(-\beta\mathcal{H}(\mathbf{D}|\mathbf{S}^o))$. Here the inverse temperature parameter is $\beta = 1$. In addition, we need to compute the maximum log-likelihood function[29], $\max_{\mathbf{D}} \log \mathcal{L}(\mathbf{D}|\mathbf{S}^o)$. This can be obtained by slowly increasing the inverse temperature parameter β in the above Monte Carlo algorithm. Note that in this way we obtain approximate values for the marginal probabilities and the maximum log-likelihood. The quality of these approximations of course depends on the computation time we spend for equilibration of the system in the Monte Carlo algorithm and the annealing process.

In practice, for a sparse D2S1 model with $N_D = 50$ diseases, $N_S = 500$ signs, and graph parameters $M_a = 50$, $M_{ab} = 100$, $k_a = k_{ab} = 150$, we run the algorithm for $\mathcal{N}_{total} = 20000$ iterations (Monte Carlo sweeps) with $\mathcal{N}_{eq} = 2000$ iterations for equilibration of the system, and extract the sample configurations after any $\mathcal{N}_{sample} = 20$ iterations. To compute the maximum log-likelihood with the annealing algorithm, we increase linearly the inverse temperature parameter β from 1 to 10 in $\mathcal{N}_{annealing} = 5000$ iterations. Altogether, computing all the necessary marginal probabilities and the objective function for a given sequence of $T = 20$ observations with the above parameters takes about 20 minutes of CPU time in a standard computer.

References

1. Kasper, D. *et al*. Harrison's Principles of Internal Medicine. 19 edition, 3000 pages, ISBN: 0071802150 (McGraw-Hill Education 2015).
2. Papadakis, M., McPhee, S. J. & Rabow, M. W. Current Medical Diagnosis and Treatment. 55 edition, LANGE CURRENT Series, 1920 pages, ISBN: 0071845097 (McGraw-Hill Education 2016).
3. Miller, R. A. & Geissbuhler, A. Clinical Diagnostic Decision Support Systems-An Overview, Page 3–34, ISBN: 978-1-4757-3903-9 (Springer New York 1999)
4. Ledley, R. S. & Lusted, L. B. Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physiciansreason. *Science* **130**, 9–21 (1959).
5. Adams, J. B. A probability model of medical reasoning and the MYCIN model. *Mathematical biosciences* **32**(1), 177–186 (1976).
6. Miller, R. A., Pople, H. E. Jr. & Myers, J. D. Internist-I, an experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine* **307**(8), 468–476 (1982).
7. Buchanan, B. G. and Shortliffe, E. H. (Eds). Rule-based expert systems (Vol. 3). Reading, MA (Addison-Wesley, 1984).
8. Miller, R., Masarie, F. E. & Myers, J. D. Quick medical reference (QMR) for diagnostic assistance. *MD computing: computers in medical practice* **3**(5), 34–48 (1985).
9. Barnett, G. O., Cimino, J. J., Hupp, J. A. & Hoffer, E. P. DXplain: an evolving diagnostic decision-support system. *JAMA* **258**(1), 67–74 (1987).
10. Spielgelharter, D. J. Probabilistic Expert Systems in Medicine. *Statistical Science* **2**, 3–44 (1987).
11. Bankowitz, R. A. *et al*. A computer-assisted medical diagnostic consultation service: implementation and prospective evaluation of a prototype. *Annals of Internal Medicine* **110**(10), 824–832 (1989).
12. Heckerman D. A tractable inference algorithm for diagnosing multiple diseases, In Machine Intelligence andPattern Recognition: Uncertainty in artificial Intelligence 5, Henrion M., Shachter R., Kanal L. N., Lemmer J. F., eds. Amsterdam North Holland Publ. Comp. 163–172 (1990).
13. Shwe, M. A. *et al*. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base. *Methods of information in Medicine* **30**(4), 241–255 (1991).
14. Heckerman, D. E. & Shortliffe, E. H. From certainty factors to belief networks. *Artificial Intelligence in Medicine* **4**(1), 35–52 (1992).
15. Nikovski, D. Constructing Bayesian networks for medical diagnosis from incomplete and partially correct statistics. *Knowledge and Data Engineering, IEEE Transactions on* **12**(4), 509–516 (2000).
16. Miller, R. A. Medical diagnostic decision support systems–past, present, and future: a threaded bibliography and brief commentary. *J Am Med Inform Assoc* **1**, 827 (1994).
17. Reggia, J. A., Nau, D. S. & Wang, P. Y. Diagnostic expert systems based on a set covering model. *Int J Man Mach Stud* **19**, 437460 (1983).
18. Berman, L. & Miller, R. A. Problem area formation as an element of computer aided diagnosis: a comparison of two strategies within quick medical reference (QMR). *Methods Inf Med* **30**, 9095 (1991).
19. Jordan, M. I. Graphical models. *Statistical Science* 140–155 (2004).
20. Murphy, K. P. Machine learning: a probabilistic perspective, (MIT press 2012).
21. Birge, J. R. & Louveaux, F. Introduction to stochastic programming. (Springer Science & Business Media 2011).
22. Altarelli, F., Braunstein, A., Ramezanpour, A. & Zecchina, R. Stochastic matching problem. *Physical review letters* **106**(19), 190601 (2011).
23. Ramezanpour, A. & Mashaghi, A. Toward First Principle Medical Diagnostics: On the Importance of Disease-Disease and Sign-Sign Interactions. *Frontiers in Physics* **5**, 32 (2017).

24. Tanaka, T. Mean-field theory of Boltzmann machine learning. *Physical Review E* **58**(2), 2302 (1998).
25. Hartmann, Alexander K., and Heiko Rieger, eds. New optimization algorithms in physics. (John Wiley & Sons 2006).
26. Mezard, M. & Montanari, A. Information, physics, and computation. (Oxford University Press 2009).
27. Ricci-Tersenghi, F. The Bethe approximation for solving the inverse Ising problem: a comparison with other inference methods. *Journal of Statistical Mechanics: Theory and Experiment* **8**, P08015 (2012).
28. Nguyen, H. C., Zecchina, R. & Berg, J. Inverse statistical problems: from the inverse Ising problem to data science. *Advances in Physics* **66**(3), 197–261 (2017).
29. Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Burlington, MA (Morgan Kaufmann 2014).
30. Press, S., Ghosh, K., Lee, J. & Dill, K. A. Principles of maximum entropy and maximum caliber in statistical physics. *Reviews of Modern Physics* **85**(3), 1115 (2013).
31. Garey, M. R. & Johnson, D. S. Computers and intractability: A guide to the theory of NP-completeness. ISBN: 0716710447 (W.H. Freeman 1979).
32. Menche, J. *et al.* Uncovering disease-disease relationships through the incomplete interactome. *Science* **347**(6224), 1257601 (2015).

Author Contributions

A.M. conceived the project and designed the study. A.M. and A.R. performed the research and analyzed the data. A.R. wrote the Monte Carlo code. A.M. and A.R. wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-20826-y.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.