A Bayesian Hierarchical Network for Combining Heterogeneous Data Sources in Medical Diagnoses

Claire Donnat   CDONNAT@UCHICAGO.EDU
Department of Statistics, University of Chicago

Nina Miolane   NMIOLANE@STANFORD.EDU
Department of Computer Science, University of California Santa Barbara

Freddy Bunbury   FBUNBURY@CARNEGIESCIENCE.EDU
Carnegie Institution for Science

Jack Kreindler   DRJACK@CHHP.COM
Centre for Health and Human Performance

Abstract

The increasingly widespread use of affordable, yet often less reliable medical data and diagnostic tools poses a new challenge for the field of Computer-Aided Diagnosis: how can we combine multiple sources of information with varying levels of precision and uncertainty to provide an informative diagnosis estimate with confidence bounds? Motivated by a concrete application in lateral flow antibody testing, we devise a Stochastic Expectation-Maximization algorithm that allows the principled integration of heterogeneous and potentially unreliable data types. Our Bayesian formalism is essential in (a) flexibly combining these heterogeneous data sources and their corresponding levels of uncertainty, (b) quantifying the degree of confidence associated with a given diagnostic, and (c) dealing with the missing values that typically plague medical data. We quantify the potential of this approach on simulated data, and showcase its practicality by deploying it on a real COVID-19 immunity study.

Keywords: Bayesian Networks; Computer-Aided Diagnostics; Affordable Healthcare; COVID-19;

1. Introduction and Related Work

Current medical diagnoses are most often based on the combination of several data inputs by medical experts, typically including (i) clinical history, interviews, and physical exams, (ii) laboratory tests, (iii) electrophysiological signals, and medical images. Advances in the machine learning (ML) community have highlighted the potential of ML to contribute to the field of Computer-Aided Diagnosis (CAD), for which we distinguish two main classes of methods.

Single modality analysis. Most recent efforts in the ML community have focused on analysing a single medical data source — called a "modality". For instance, the parsing of electronic health records (EHR) for clinical history, patients' interviews, and physical exams has shown huge potential for the diagnosis of a broad range of diseases, from coronary artery disease to rheumatoid arthritis Harrell

et al. (1984); Kurt et al. (2008); Carroll et al. (2011); Hippisley-Cox and Coupland (2013); Rahimian et al. (2018). ML algorithms have been developed to process various types of laboratory results, from urine steroid profiles to gene expression data Huang and Wu (2018); Wilkes et al. (2018); Demirci et al. (2016); Agrahari et al. (2018). Others have shown success in automatically processing and classifying medical signals such as electrocardiograms (ECG) Acharya et al. (2017); Kiranyaz et al. (2016); Rad et al. (2017); Pławiak (2018), or electroencephalograms (EEG) Richhariya and Tanveer (2018). Finally, a large body of work has focused on medical imaging analysis across multiple tasks, such as automatic extraction of diagnostic features Thomaz et al. (2017); Bar et al. (2018), segmentation of anatomical structures Moeskops et al. (2016); Li et al. (2015); Shelhamer et al. (2017); Milletari et al. (2016); Dolz et al. (2018); Christ et al. (1919), or direct diagnosis through image classification Anthimopoulos et al. (2016); Miki et al. (2017).

While these works achieve record-breaking diagnostic accuracy, they often rely on supervised learning approaches – requiring the diagnosis ground-truth to be available during training – and on the acquisition of large datasets. Furthermore, they focus solely on a single type of data input (EEGs, scans, etc.) – often acquired by clinicians using specialized, highly accurate equipment – and do not harvest the potentially rich and complementary sources of information provided by alternative medical modalities. By contrast, the development of at-home diagnostic tests (lateral flow assays, questionnaire data for disease screening, mobile health apps, etc.) paves the way for more accessible healthcare, which could be significantly improved by establishing a method to combine these often noisy data sources. Given the uncertainty exhibited by these new inputs, it also becomes indispens-

able to pair the provision of a diagnosis with a notion of a confidence interval.

Multiple modality integration. Interestingly, diagnosis studies fusing different data sources seemed more prominent in the first years of CAD. Bayesian networks Neal (1993) – or belief networks – have been used as a crucial decision tool for automatic diagnosis. Such networks provide a medically-grounded and interpretable statistical framework that integrates the probabilities of contracting the disease given a patient's clinical history, and the probability of developing symptoms or observing positive test results in the presence or absence of the disease. The advantages of these Bayesian networks are that they are (a) fully unsupervised and do not require ground-truth information, and (b) able to provide meaningful results even in low sample regimes.

Bayesian networks have thus been implemented in a variety of contexts to integrate clinical data and laboratory results, and diagnose conditions ranging from pyloric stenosis to acute appendicitis Alvarez et al. (2006); Gevaert et al. (2006); González-López et al. (2016); Sangamuang et al. (2018). They have also been used to combine clinical data with medical images, and subsequently applied to assess venous thromboembolism Kline et al. (1984), community-acquired pneumonia Aronsky and Haug (1998), or to predict tumors Wang et al. (1999); Sneha and Agrawal (2013). As early as 1994, full integration of clinical data, laboratory results and imaging features was performed to diagnose gallbladder disease Haddawy et al. (1994).

However, contrary to our proposed setting, the uncertainty that these Bayesian networks integrate is typically known and controlled. The medical conditions that they study are well characterized by a set of specific questions and physiological exams (e.g. projectile vomiting, potassium levels and ultrasounds in the case of Pyloric stenosis). Not only do these inputs provide a very strong signal for

the diagnosis, but the uncertainty arising from the different modalities can often be reliably informed by a test manufacturer, extensive medical research or prior ML studies. The uncertainty of a diagnostic method, however, is not always well specified: whether it be (i) a physician's assessment of the symptoms associated with a novel or rare disease, (ii) predictions of an ML algorithm whose accuracy has not been fully characterized outside of a curated research environment or reference datasets, or (iii) biological tests whose sensitivity still has to be determined.

Multiple noisy modality integration: a COVID-19 case study. To motivate this paper and further understand the limitations of previous approaches, let us consider a particular use case: COVID-19 antibody testing. Lateral flow immunoassay (LFA) antibody tests for COVID-19 are one of the most manageable, affordable and easily deployable options to allow at-home testing for large population and provide assessments of our past exposure and immunity. Yet, studies have shown that the sensitivity of these tests remains highly variable and highly contingent on the time of testing. The successful deployment of LFAs thus depends on their augmentation with additional data inputs, such as user-specific risk factors and self-reported symptoms. The provision of confidence scores is essential to flag potential false negative or positive tests (requiring re-testing) and to assess local prevalence levels – both pivotal for researchers and policymakers in the context of a pandemic.

Contributions and outline. This applications paper is geared towards the practical integration of noisy sources of information for CAD. Our contribution is two-fold. From a methods perspective, we account for the variability of the inputs by devising a two-level Bayesian hierarchical model. In contrast to existing Bayesian methods for CAD, our model is deeper, and trained using a

Stochastic Expectation Maximization (EM) algorithm Celeux et al. (1995, 2001); Nielsen et al. (2000). The Stochastic EM overcomes the limitations of its non-stochastic counterpart Celeux et al. (1995), that is (a) its sensitivity to the starting position, (b) its potentially slow convergence rate, and (c) its possible convergence to a saddle position instead of a local maximum. From an applications perspective, we gear this algorithm to enhance at-home LFA testing. In particular, we wish to (a) quantify the benefit of multimodal data integration when the diagnostics are uncertain, and (b) show how our method can benefit medical experts or researchers in real life. We also provide a way to incorporate incomplete data entries in training — which typically plague self-reported data. Section 3 presents the Bayesian model for multimodal integration and the Stochastic EM algorithm that performs principled and scalable inference. Section 5 presents extensive tests of our model on simulated datasets. Section 6 details the results obtained on the Covid-19 dataset and the impact of our method for affordable and reliable at-home test kits.

2. Covid-19 Dataset

By way of clarifying the challenges that our algorithm proposes to overcome, we present here the COVID Clearance Remote Recovery & Antibody Monitoring Platform study [1], which motivated our approach. The purpose of this study is to track the evolution of the immunity of a cohort of participants in the UK with various COVID-19 exposure risks. This paper focuses on a subset of 261 participants, for whom we have both questionnaire information and LFA test results.

Home-testing Immunoassay Data. Participants are issued packs of home-testing kits with written instructions. These kits identify Immunoglobulin M (IgM) and/or Immunoglobulin G (IgG) specific to SARS-CoV2

---

1. COV-CLEAR, www.cov-clear.com

in blood samples. Pictures and typical examples of LFA test results are provided in the supplementary materials. Participants self-report a positive result if IgM and/or IgG are detected by the test in addition to a positive control band. The COV-CLEAR study aims to address questions relating to (a) the quantification of the robustness of the antibody response, and (b) the durability of this response. However, while scalable and affordable, LFA tests suffer from vastly varying levels of sensitivity — with a sensitivity as low as 70% on asymptomatic individuals.

Clinical Data: Questionnaire. Additionally, participants are asked to answer a questionnaire, ideally before knowing the result of their LFA test. The form consists of questions related to $K = 14$ exhibited symptoms (fever, cough, runny nose, etc.) and $M = 2$ subject-specific risk factors (household size and proportion of members with a suspected or confirmed history of COVID-19). The empirical distributions of the symptoms and the risk factors in this cohort are provided in the supplementary materials.

3. Multimodal Data Integration

In this section, we describe our principled integration of noisy diagnostic test results, with additional clinical data such as symptom data and subject-specific risk factors. Our approach applies to general heterogeneous medical data where the outputs are binary. The latter could be either self-reported answers to questionnaires, clinician-reported physiological exams, outputs of a diagnostic based on an image, abnormalities of laboratory results, etc., making this a widespread and general setting. Thus, while we apply this method to antibody testing, it could be relevant to any medical diagnostic with binary inputs.

Denote D the true state of an individual (healthy/sick), T the outcome of the noisy diagnostic test (positive/negative), S the symptomaticity (symptomatic/asymptomatic), X the symptoms and Y the risks factors. The underlying assumption is that given a true diagnosis state D, the symptoms X and the diagnostic test outcome T are independent. In other words, the probability of the diagnostic test being a false negative $\mathbb{P}[T = 0|D = 1]$ is independent of the symptoms of a truly infected individual $\mathbb{P}[X|D = 1]$. Similarly, given a state D, the test outcome T and the exhibited symptoms X are independent of the risk factors Y. We define:

- $\mathbb{P}(D = 1|Y) = \pi_\beta(Y)$ the probability of being ill given risk factors Y,
- $\mathbb{P}[(T = 1|D = 1) = x$ the sensitivity of the diagnostic test,
- $\mathbb{P}[T = 0|D = 0] = 1 - y$ the specificity of the diagnostic test, *i.e.* $y = \mathbb{P}[T = 1|D = 0]$,
- $\mathbb{P}(S = 1|D = 0) = p_0$ the probability of being symptomatic whilst not having been infected by that specific disease (the symptoms could be due to another illness),
- $\mathbb{P}(S = 1|D = 1) = p_1$ the probability of having been symptomatic upon infection,
- $\mathbb{P}(X_k = 1|S = 1, D = 0) = s_{0k}$ the probability of exhibiting symptom k when not infected,
- $\mathbb{P}(X_k = 1|S = 1, D = 1) = s_{1k}$ the probability of exhibiting symptom k upon infection.

In the above, the diagnostic test may be any medical test (e.g. LFA antibody tests, or output of a medical imaging classification algorithm). We have here split our binary variables into two classes according to their variability: (a) the test T, for which we have provisional estimates for the sensitivity and specificity (given by the manufacturer) but which we are still uncertain, and (b) the symptoms X, which carry additional uncertainty in that neither of them is extremely specific to COVID-19 nor is their prevalence amongst COVID cohorts very well established; Values for (x, y), p_0, p_1, s_0, s_1 or β may be published, but challenged by complementary research or

field experience, or may be completely unavailable. We thus need to account for these inputs' variability and for the relative importance of different risk factors β. We propose using a hierarchical Bayesian network (see Fig. 9) that is consequently deeper than the ones traditionally implemented in the CAD literature Alvarez et al. (2006); Gevaert et al. (2006); Sakai et al. (2007); González-López et al. (2016); Sangamuang et al. (2018).

The uncertainty in the test sensitivity x and specificity $1 - y$ are expressed through Beta priors on x and y with parameters (α_x, β_x) and (α_y, β_y), see Fig. 9. Similarly, the uncertainty in the probabilities of symptomaticity and of the different symptoms are expressed via Beta priors with respective parameters $(\alpha_{p_0}, \beta_{p_0}), (\alpha_{p_1}, \beta_{p_1}), \{(\alpha_{s_{0k}}, \beta_{s_{0k}}), (\alpha_{s_{1k}}, \beta_{s_{1k}})\}_k$, see Fig. 9. When published estimates exist for x, y, p_0, p_1, s_0 and s_1 (e.g. as provided by the LFA test manufacturer), we match the moments of the Beta priors with those of the published distributions. If this information is unavailable, we use the non-informative Beta prior with parameters $(0.5, 0.5)$. In the COVID-19 example, we implement a non-informative prior for the probabilities of symptomaticity and symptoms, as the etiology of the disease remains unknown. These priors are then updated during learning as we aggregate the information from the observed and imputed variables.

Lastly, we model the probability $\pi_\beta(Y)$ of contracting the disease with a logistic regression on the M risk factors Y. In the case of the COVID-19 dataset, these include county data, profession, size of household, etc. The coefficients $\beta = \{\beta_m\}_m$ weight the relative importance of each factor $Y_m, m = 1...M$. We express our uncertainty on the possible impact of a given factor m by introducing a Gaussian prior: $\beta \sim N(0, \sigma_\beta^2)$, see Fig. 9.

4. Inference via Stochastic EM

At the subject level, we seek to compute the posterior distribution of the true diagnosis D_i, informed by the integration of the observed variables T_i, S_i, X_i, Y_i. This posterior yields an estimated diagnosis, together with a credible interval that expresses the confidence associated with our prediction. In the case of COVID-19, this provides individual estimates of each citizen's immunity, which may inform their social behavior as lockdown is eased.

At the global level, we wish to learn (i) the distributions of the sensitivity x and specificity y of the diagnostic test; (ii) the distributions of the probabilities p_0, p_1 of being symptomatic, and s_0, s_1 of exhibiting specific symptoms, within a population of interest; and (iii) the distribution of the impact β of the risk factors for contracting the disease. In the context of the COVID-19 pandemic, such aggregated figures are pivotal to understanding the dynamics of the disease and to implementing appropriate crisis policy.

Since the true diagnoses are hidden variables, the Expectation-Maximization algorithm Dempster et al. (1977) is an appealing method to perform inference in our Bayesian model; hence to achieve the aforementioned objectives. However, the EM algorithm requires the computation of an expectation over the posterior of the hidden variables, which may be intractable depending on the probability distributions defined in the model. To allow for flexibility in terms of model's design within our multimodal data integration framework, we opt for the Stochastic EM algorithm (StEM) Celeux and Diebolt (1988). StEM effectively estimates the conditional expectation in the EM using the "Stochastic Imputation Principle", approximating the expectation by sampling once from the underlying distribution. This method allows us to carry out inference with priors that are not necessarily the Beta distributions implemented in our experiments — thus providing additional flexibility in modelling real data. StEM is also more robust, being less dependent on the parameters' initialization than
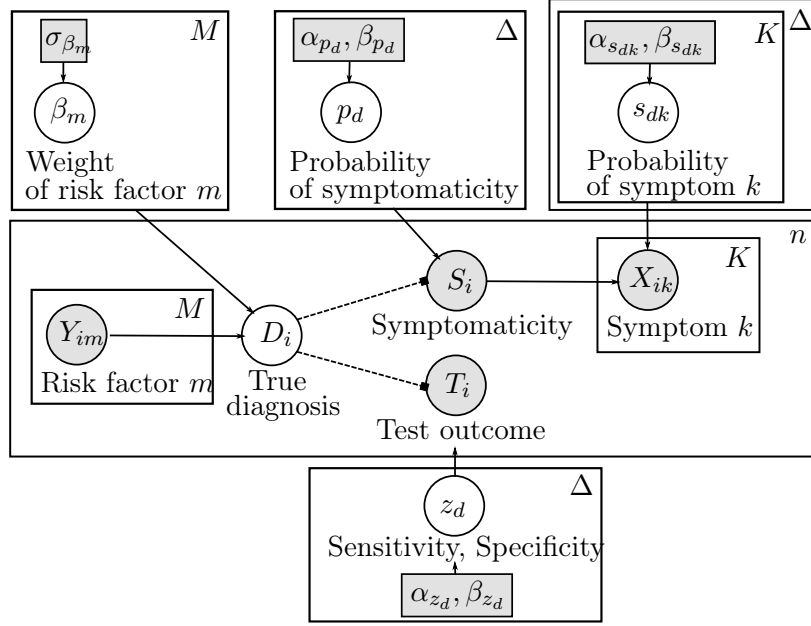
Figure 1: Hierarchial bayesian network integrating the accuracy uncertainty on the data sources, to estimate the true diagnosis D. The index $k = 1...K$ represents symptom k, while $m = 1...M$ represents risk factor m. The shaded cells represent observed variables or known hyper-parameters. The dashed lines represent switches.

EM – a definite advantage given our uncertain framework. Finally, StEM shows better asymptotic behavior: unlike the EM, StEM always leads to maximization of a complete data log-likelihood in the M-step Nielsen (2000).

4.1. Stochastic E-step at iteration $(j+1)$

Starting iteration $j + 1$, the current estimates of the model's parameters are $\theta^{(j)}$. The stochastic E-step computes the posterior distribution $\mathbb{P}(D_i|T_i, S_i, X_i, Y_i, \theta^{(j)})$ of the hidden variable D_i. We then sample a diagnosis $\widehat{D}_i$ from this posterior.

Proposition 1 *The odds of the posterior of the hidden variable D_i at iteration $(j+1)$ writes:*

$$\frac{\mathbb{P}(D_i = 1|T_i, S_i, X_i, Y_i, \theta^{(j)})}{\mathbb{P}(D_i = 0|T_i, S_i, X_i, Y_i, \theta^{(j)})}$$

$$= \frac{x^{(j)T_i}(1 - x^{(j)})^{(1-T_i)}s_1^{(j)S_i X_i}(1 - s_1^{(j)})^{S_i(1-X_i)}}{y^{(j)T_i}(1 - y^{(j)})^{(1-T_i)}s_0^{(j)S_i X_i}(1 - s_0^{(j)})^{S_i(1-X_i)}}$$

$$\times \frac{\pi(Y_i, \beta^{(j)}) \times p_1^{S_i}(1 - p_1)^{(1-S_i)}}{(1 - \pi(Y_i, \beta^{(j)})) \times p_0^{S_i}(1 - p_0)^{(1-S_i)}}. \tag{1}$$

4.2. M-step at iteration $(j+1)$

The following proposition shows the updates in the model's parameters at iteration $(j+1)$, performed in the M-step.

Proposition 2

The parameters updates write:

$$x^{(j+1)} = \frac{\sum_{i=1}^{n} T_i \widehat{D_i} + \alpha_x - 1}{\sum_{i=1}^{n} \widehat{D_i} + (\alpha_x + \beta_x) - 2}$$

$$y^{(j+1)} = \frac{\sum_{i=1}^{n} T_i(1 - \widehat{D_i}) + \alpha_y - 1}{\sum_{i=1}^{n}(1 - \widehat{D_i}) + (\alpha_y + \beta_y) - 2}$$

$$p_0^{(j+1)} = \frac{\sum_{i=1}^{n}(1 - \widehat{D_i})S_i + \alpha_{p_0} - 1}{\sum_{i=1}^{n}(1 - \widehat{D_i}) + (\alpha_{p_0} + \beta_{p_0}) - 2}$$

$$p_1^{(j+1)} = \frac{\sum_{i=1}^{n} \widehat{D_i}S_i + \alpha_{p_1} - 1}{\sum_{i=1}^{n} \widehat{D_i} + (\alpha_{p_1} + \beta_{p_1}) - 2}$$

$$s_0^{(j+1)} = \frac{\sum_{i:S_i=1}(1 - \widehat{D_i})X_i + \alpha_{s_0} - 1}{\sum_{i:S_i=1}^{n}(1 - \widehat{D_i}) + (\alpha_{s_0} + \beta_{s_0}) - 2}$$

$$s_1^{(j+1)} = \frac{\sum_{i:S_i=1} \widehat{D_i}X_i + \alpha_{s_1} - 1}{\sum_{i=1,s.t.\ S_i=1}^{n} \widehat{D_i} + (\alpha_{s_1} + \beta_{s_1}) - 2}$$

$$\beta^{(j+1)} = argmin_\beta \sum_{i=1}^{n} \frac{||\widehat{D_i} - g(Y_i\beta)||^2}{2\sigma^2} + \frac{||\beta||^2}{2\sigma_\beta^2},$$

where the minimization on β is performed through Newton-Raphson descent.

Algorithm Complexity. Our algorithm only relies on simple sequential updates of the distribution parameters, as highlighted in Prop. 3 and 4. Denoting by L the number of such parameters, and N the number of samples, the updates for X and T only involve matrix multiplications of the form $\widehat{D}^T X$, while those for β involve matrix multiplications of size $O(M^2N)$ — yielding a complexity of $O(LN)$. Prior updates rely solely on element-wise operations on the log-odds, with $O(N)$ complexity. Denoting as B the maximum number of steps, the complexity is $O(BLN)$, thus linear in N.

5. Validation on Synthetic Data

Since our approach is unsupervised, we begin by validating it on synthetic datasets where the ground truth is known and controlled — allowing us to characterize the performance of the algorithm, and to showcase the strength of combining multiple noisy modalities.

We assume the same generative model as in Fig. 9 and generate synthetic data for various pairs of values of sensitivity and specificity ranging from 60% to 99% . In each case, we simulate $N = 100$ tuples of variables $((Y_i, X_i, T_i)_{i=1}^n$, for $n \in \{100, 200, 500, 1000\}$ participants. We also vary the noise level $\sigma \in \{0.1, 0.5, 1.0\}$ for the prior on D in Fig. 9. To mimic our COVID data, we simulate $K = 14$ symptoms and $M = 2$ risk factors, for which we randomly select the parameters[2].

Improving upon the sole test T. Fig. 10 (A) and (B) show the Stochastic EM's raw accuracy and accuracy improvement over the sole test result T. Note that this difference is always positive, highlighting that our method only improves upon single diagnostic inputs — even when one input is more reliable than any of the others. As expected, the most substantial improvements upon T are observed when the test specificity and/or sensitivity are low. For instance, for sensitivity and specificity of 70%, our method provides an $86.6 \pm (2.5)\%$ accuracy for the diagnosis – or equivalently, a 16% gain over T. We further quantify the algorithm's performance in Table 1 of the supplementary materials, for values of the specificity close to those observed on the field.

Benchmarking against other models. We now benchmark our algorithm against other approaches:

- *Vanilla Classifier:* using both the context Y and symptoms X as inputs, we fit a logistic regression to the test labels T. We choose the regularization parameter using 10-fold cross-validation, and compute confidence intervals for the log-probability of $(D|X, T)$ by bootstrapping.

- *Data-Agnostic EM:* we implement a deterministic version of EM, providing uninformative priors for the parameter (thereby reflecting our absence of knowledge of the

---

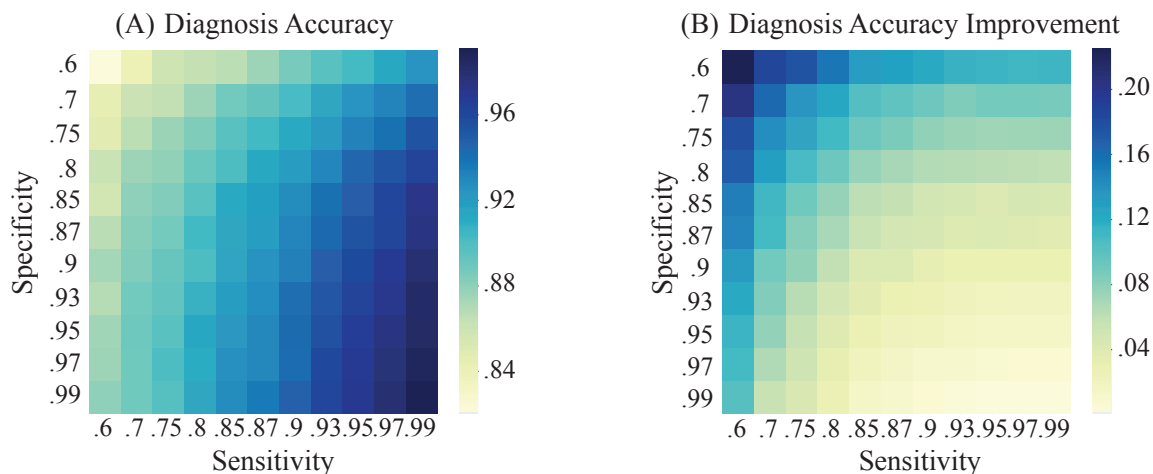2. All of these parameters are provided with the code in the supplementary materials.

Figure 2: Performance of the SEM algorithm for n=300 samples, $\sigma = 0.5$, and varying levels of specificity and sensitivity. **(A)** Raw accuracy of the labels imputed via StEM. **(B)** Difference in accuracy between the StEM imputed diagnostics and the observed test T.

truth), which are not updated — i.e, an "uninformed" equivalent of the belief networks found in the literature.

- *Data-Informed EM*: similar to the Data-Agnostic EM, but choosing the priors (which then remain fixed) based on the empirical data.

The results are shown in Fig. 3(C), and in the supplementary materials. We highlight the superiority of our deeper and more adaptive hierarchical model, yielding improvements (for a reasonable tuple of specificity and sensitivity of 80%) of up to 4% over the Data-Informed EM, 8% over the Data-Agnostic one, and 9% over the Vanilla classifier.

Assessing the robustness of the Stochastic EM. For a fixed specificity 80%, Figure 3 shows the accuracy of StEM for different values of σ (A) or as the number of samples increases (B). These axes of variation seem to have little impact on the performance, providing evidence for the algorithm's robustness, especially with respect to low-sample size.

Convergence. Finally, we examine the distribution of the relative difference between recovered coefficients and ground truth (expressed as a percentage of the ground-truth

value). The plots, provided in the supplementary materials, show deviations that are within a few percentages of the true value of the coefficients – thus highlighting the ability of the model to converge to the ground-truth parameters, and making it relevant from a medical perspective to characterize the disease's etiology.

6. Results on Real Data

We turn to the real dataset of $n = 261$ participants described in Section 2. The purpose of this section is to show how our algorithm can be practically applied to process heterogeneous data types and inform participants and researchers alike. At the individual level, we provide each user with (a) the confirmation of the diagnostic, and (b) confidence intervals associated with the uncertainty to shed more light on the uncertainty associated with the diagnostic. At the global level, we provide policymakers with (c) aggregated analysis of the herd immunity, and associated measures of uncertainty. For the purpose of clarity, we showcase here a few figures and properties. We defer to the appendix the discussion of the pre-processing and the potential sources
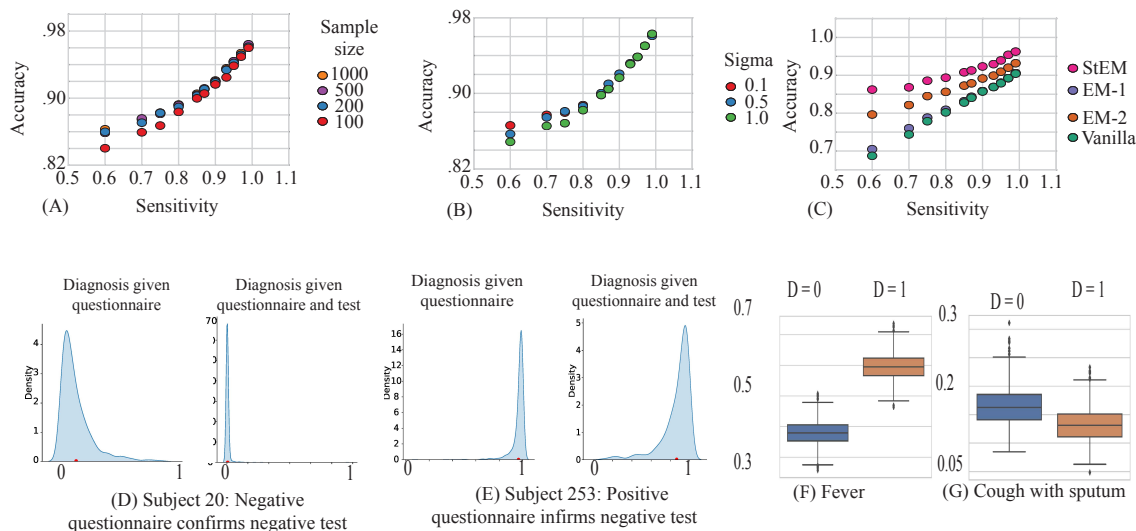
Figure 3: Performance of the StEM for a fixed specificity of 80%, (A) as n increases ($\sigma = 0.5$);(B) as the variance increases ($n = 300$) and (C) Accuracy (w.r.t. benchmarks). (D-E) Posterior diagnosis distributions for two subjects. Left: posterior of the diagnosis, given the questionnaire; right: posterior, given the questionnaire and reported test result. (F-G): Posteriors probabilities of specific symptoms, for subjects with $D = 0$ or $D = 1$.

of bias (and how to correct for them) of this particular dataset.

At the subject-level. We present two examples, where our algorithm either confirms or invalidates the result of the test – thereby allowing for the potential flagging of false negatives (Fig. 3D-E). The second example showcases an instance where the questionnaire and the test disagree. Subject 253 is a user that registers a negative test, while exhibiting a wide number of known COVID symptoms (dry cough, shortness of breath, fever), but took the test less than 10 days after his illness (when antibody levels are often just starting to increase). While the posterior does not reclassify the subjects' diagnosis, the confidence interval associated with the prediction of immunity reflects the uncertainty associated with this case, and flags it as a potential false negative. Similar plots on additional symptoms (which can be non-binary) are provided in the supplementary materials

and an interactive dashboard showcasing the algorithm and the effect of the different symptoms can be found at the following link[3] — potentially allowing users to further inform and trust their antibody test results.

At the global level. This framework paves the way for principled inference for the disease and the population of interest. One could indeed use these probabilities in order to gauge the risk associated to a given community, or to inform testing policies. The posterior distributions of the model's parameters also shed light on the actual accuracy of the LFA tests on our population. Our results provide information regarding the most prevalent symptoms for COVID-19 (Fig. 3F,G). For instance, among symptomatic participants, the probability of exhibiting fever is higher for infected subjects (see Fig. 3 (B)) while cough with sputum does not seem to be associated with COVID-19,

---

3. https://homecovidtests.shinyapps.io/ COVID-app/

being probably the sign of another infection. **Dealing with missing data.** Finally, one of the strengths of the Bayesian formalism lies in its ability to impute missing information and leverage strength from partially observable inputs — a crucial element for self-reported data, and the derivations for which, in our case, can be found in the appendix.

7. Conclusion

This applications paper provides a statistical framework for the integration of noisy diagnostic test results, self-reported symptoms, and demographic variables. Compared to previous approaches, our work is more amenable to the handling of the different inputs' uncertainty. While we have focused here on binary symptoms, this adaptive Bayesian framework, together with the flexibility provided by the Stochastic EM algorithm, pave the way for the integration of other continuous-valued inputs (index of the severity of the disease, pain, etc. — see appendix). The robustness of the Stochastic EM is crucial in ensuring the reliability of the parameters given the medical stakes, while allowing diagnoses even in the presence of incomplete data.

References

U. Rajendra Acharya, Hamido Fujita, Oh Shu Lih, Yuki Hagiwara, Jen Hong Tan, and Muhammad Adam. Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network. *Information Sciences*, 405:81–90, 2017. ISSN 00200255. doi: 10.1016/j.ins.2017.04.012. URL http://dx.doi.org/10.1016/j.ins.2017.04.012.

Rupesh Agrahari, Amir Foroushani, T. Roderick Docking, Linda Chang, Gerben Duns, Monika Hudoba, Aly Karsan, and Habil Zare. Applications of Bayesian network models in predicting types of hematological malignancies. *Scientific Reports*, 8(1): 1–12, 2018. ISSN 20452322. doi: 10.1038/ s41598-018-24758-5. URL http://dx.doi.org/10.1038/s41598-018-24758-5.

Sonia M. Alvarez, Beverly A. Poelstra, and Randall S. Burd. Evaluation of a Bayesian decision network for diagnosing pyloric stenosis. *Journal of Pediatric Surgery*, 41 (1):155–161, 2006. ISSN 00223468. doi: 10.1016/j.jpedsurg.2005.10.019.

Marios Anthimopoulos, Stergios Christodoulidis, Lukas Ebner, Andreas Christe, and Stavroula Mougiakakou. Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network. *IEEE Transactions on Medical Imaging*, 35(5): 1207–1216, 2016. ISSN 1558254X. doi: 10.1109/TMI.2016.2535865.

D. Aronsky and P. J. Haug. Diagnosing community-acquired pneumonia with a Bayesian network. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, (June 1997):632–636, 1998. ISSN 1531605X.

Yaniv Bar, Idit Diamant, Lior Wolf, Sivan Lieberman, Eli Konen, and Hayit Greenspan. Chest pathology identification using deep feature selection with non-medical training. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization*, 6(3):259–263, 2018. ISSN 21681171. doi: 10.1080/21681163.2016.1138324. URL http://dx.doi.org/10.1080/21681163.2016.1138324.

Robert J. Carroll, Anne E. Eyler, and Joshua C. Denny. Naïve Electronic Health Record phenotype identification for Rheumatoid arthritis. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2011:189–196, 2011. ISSN 1942597X.

Gilles Celeux and Jean Diebolt. A random imputation principle : the stochastic EM algorithm. Technical report, 1988.

Gilles Celeux, Didier Chauveau, and Jean Diebolt. On stochastic versions of the em algorithm. 1995.

Gilles Celeux, Didier Chauveau, and Jean Diebolt. On stochastic versions of the em algorithm. *Biometrika*, 88(1):281–286, 2001. ISSN 00063444. doi: 10.1093/biomet/88.1. 281.

Patrick Ferdinand Christ, Mohamed Ezzeldin A. Elshaer, Florian Ettlinger, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfle, Marco Armbruster, Felix Hofmann, Melvin D'Anastasi, Wieland H. Sommer, Seyed-Ahmad Ahmadi, and Bjoern H. Menze. Automatic Liver and Lesion Segmentation in CT Using Cascaded Fully Convolutional Neural Networks and 3D Conditional Random Fields. *Urologic and Cutaneous Review*, 23 (8):435–460, 1919. ISSN 10636919. doi: 10.1007/978-3-319-46723-8.

Ferhat Demirci, Pinar Akan, Tuncay Kume, Ali Riza Sisman, Zubeyde Erbayraktar, and Suleyman Sevinc. Artificial neural network approach in laboratory test reporting: Learning algorithms. *American Journal of Clinical Pathology*, 146(2):227–237, 2016. ISSN 19437722. doi: 10.1093/ajcp/aqw104.

Arthur Dempster, Natalie Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

Jose Dolz, Christian Desrosiers, and Ismail Ben Ayed. 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. *NeuroImage*, 170(April 2017):456–470, 2018. ISSN

10959572. doi: 10.1016/j.neuroimage.2017. 04.039. URL https://doi.org/10.1016/ j.neuroimage.2017.04.039.

Olivier Gevaert, Frank De Smet, Dirk Timmerman, Yves Moreau, and Bart De Moor. Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. *Bioinformatics*, 22(14), 2006. ISSN 13674803. doi: 10.1093/bioinformatics/btl230.

J. J. González-López, M. García-Aparicio, D. Sánchez-Ponce, N. Muñoz-Sanz, N. Fernandez-Ledo, P. Beneyto, and M. C. Westcott. Development and validation of a Bayesian network for the differential diagnosis of anterior uveitis. *Eye (Basingstoke)*, 30(6):865–872, 2016. ISSN 14765454. doi: 10.1038/eye.2016.64.

Peter Haddawy, Charles E. Kahn, and Manonton Butarbutar. A Bayesian network model for radiological diagnosis and procedure selection: Work-up of suspected gallbladder disease. *Medical Physics*, 21(7):1185–1192, 1994. ISSN NA. doi: 10.1118/1.597400.

Frank E. Harrell, Kerry L. Lee, Robert M. Califf, David B. Pryor, and Robert A. Rosati. Regression modelling strategies for improved prognostic prediction. *Statistics in Medicine*, 3(2):143–152, 1984. ISSN 10970258. doi: 10.1002/sim.4780030207.

Julia Hippisley-Cox and Carol Coupland. Predicting risk of emergency admission to hospital using primary care data: Derivation and validation of QAdmissions score. *BMJ Open*, 3(8), 2013. ISSN 20446055. doi: 10.1136/bmjopen-2013-003482.

Lei Huang and Tong Wu. Novel neural network application for bacterial colony classification. *Theoretical Biology and Medical Modelling*, 15(1):1–16, 2018. ISSN 17424682. doi: 10.1186/s12976-018-0093-x.

Serkan Kiranyaz, Turker Ince, and Moncef Gabbouj. Real-Time Patient-Specific ECG Classification by 1-D Convolutional Neural Networks. *IEEE Transactions on Biomedical Engineering*, 63(3):664–675, 2016. ISSN 15582531. doi: 10.1109/TBME.2015.2468589.

JA Kline, AJ Novobilski, C Kabrhel, and DM Courtney. Derivation and Validation of a Bayesian Network to Predict Pretest Probability of Venous Thromboembolism. *Journal of Periodontology*, 55(5):306–313, 1984. ISSN 0022-3492. doi: 10.1902/jop.1984.55.5.306.

Imran Kurt, Mevlut Ture, and A. Turhan Kurum. Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications*, 34(1):366–374, 2008. ISSN 09574174. doi: 10.1016/j.eswa.2006.09.004.

Wen Li, Fucang Jia, and Qingmao Hu. Automatic Segmentation of Liver Tumor in CT Images with Deep Convolutional Neural Networks. *Journal of Computer and Communications*, 03(11):146–151, 2015. ISSN 2327-5219. doi: 10.4236/jcc.2015.311023.

Yuma Miki, Chisako Muramatsu, Tatsuro Hayashi, Xiangrong Zhou, Takeshi Hara, Akitoshi Katsumata, and Hiroshi Fujita. Classification of teeth in cone-beam CT using deep convolutional neural network. *Computers in Biology and Medicine*, 80:24–29, 2017. ISSN 18790534. doi: 10.1016/j.compbiomed.2016.11.003. URL http://dx.doi.org/10.1016/j.compbiomed.2016.11.003.

Fausto Milletari, Nassir Navab, and Seyed Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016*, pages 565–571, 2016. doi: 10.1109/3DV.2016.79.

Pim Moeskops, Max A. Viergever, Adrienne M. Mendrik, Linda S. De Vries, Manon J.N.L. Benders, and Ivana Isgum. Automatic Segmentation of MR Brain Images with a Convolutional Neural Network. *IEEE Transactions on Medical Imaging*, 35 (5):1252–1261, 2016. ISSN 1558254X. doi: 10.1109/TMI.2016.2548501.

Radford M Neal. *Probabilistic Inference Using Markov Chain Monte Carlo Methods*. Number September. 1993. ISBN 1095-9572 (Electronic)\r1053-8119 (Linking). doi: 10.1016/j.neuroimage.2009.01.023.

Søren Feodor Nielsen et al. The stochastic em algorithm: estimation and asymptotic results. *Bernoulli*, 6(3):457–489, 2000.

Søren Feodor Nielsen. The stochastic EM algorithm: Estimation and asymptotic results. *Bernoulli*, 6(3):457–489, 2000. ISSN 13507265. doi: 10.2307/3318671.

Paweł Pławiak. Novel methodology of cardiac health recognition based on ECG signals and evolutionary-neural system. *Expert Systems with Applications*, 92:334–349, 2018. ISSN 09574174. doi: 10.1016/j.eswa.2017.09.022.

Ali Bahrami Rad, Trygve Eftestol, Kjersti Engan, Unai Irusta, Jan Terje Kvaloy, Jo Kramer-Johansen, Lars Wik, and Aggelos K. Katsaggelos. ECG-Based classification of resuscitation cardiac rhythms for retrospective data analysis. *IEEE Transactions on Biomedical Engineering*, 64(10): 2411–2418, 2017. ISSN 15582531. doi: 10.1109/TBME.2017.2688380.

Fatemeh Rahimian, Gholamreza Salimi-Khorshidi, Amir H. Payberah, Jenny Tran,

Roberto Ayala Solares, Francesca Raimondi, Milad Nazarzadeh, Dexter Canoy, and Kazem Rahimi. Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records. *PLoS Medicine*, 15(11):1–18, 2018. ISSN 15491676. doi: 10.1371/journal.pmed.1002695.

B. Richhariya and M. Tanveer. EEG signal classification using universum support vector machine. *Expert Systems with Applications*, 106:169–182, 2018. ISSN 09574174. doi: 10.1016/j.eswa.2018.03.053. URL https://doi.org/10.1016/j.eswa.2018.03.053.

S. Sakai, K. Kobayashi, J. Nakamura, S. Toyabe, and Kohei Akazawa. Accuracy in the diagnostic prediction of acute appendicitis based on the Bayesian network model. *Methods of Information in Medicine*, 46 (6):723–726, 2007. ISSN 00261270. doi: 10.3414/ME9066.

Chaitawat Sangamuang, Peter Haddawy, Viravarn Luvira, Watcharapong Piyaphanee, Sopon Iamsirithaworn, and Saranath Lawpoolsri. Accuracy of dengue clinical diagnosis with and without NS1 antigen rapid test: Comparison between human and Bayesian network model decision. *PLoS Neglected Tropical Diseases*, 12(6):1–14, 2018. ISSN 19352735. doi: 10.1371/journal.pntd.0006573.

Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640–651, 2017. ISSN 01628828. doi: 10.1109/TPAMI.2016.2572683.

Sweta Sneha and Rajeev Agrawal. Towards enhanced accuracy in medical diagnostics - A technique utilizing statistical and clinical data analysis in the context of ultrasound images. *Proceedings of the Annual Hawaii International Conference on System Sciences*, pages 2408–2415, 2013. ISSN 15301605. doi: 10.1109/HICSS.2013.566.

Ricardo L. Thomaz, Pedro C. Carneiro, and Ana C. Patrocinio. Feature extraction using convolutional neural network for classifying breast density in mammographic images. *Medical Imaging 2017: Computer-Aided Diagnosis*, 10134:101342M, 2017. ISSN 16057422. doi: 10.1117/12.2254633.

Xiao Hui Wang, Bin Zheng, Walter F. Good, Jill L. King, and Yuan Hsiang Chang. Computer-assisted diagnosis of breast cancer using a data-driven Bayesian belief network. *International Journal of Medical Informatics*, 54(2):115–126, 1999. ISSN 13865056. doi: 10.1016/S1386-5056(98)00174-9.

Edmund H. Wilkes, Gill Rumsby, and Gary M. Woodward. Using machine learning to aid the interpretation of urine steroid profiles. *Clinical Chemistry*, 64(11):1586–1595, 2018. ISSN 15308561. doi: 10.1373/clinchem.2018.292201.

Appendix A. COVID-19 Datasets

This section provides additional details on the COV-CLEAR dataset [4]. Figure 4 shows the sampling distributions of selected symptoms and risk factors in the population of interest – grouped by the result of the lateral flow immunoassay (LFA) test. Figure 5 illustrates the LFA test results as observed by a given subject. Participants self-report a positive result if IgM and/or IgG are detected by the test. The marker "C" stands for "Control".
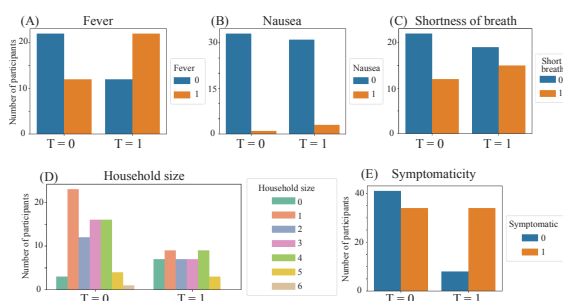

Figure 4: (A-C) Distributions of selected symptoms in the COV-CLEAR dataset. (D) Distribution of household size, a COVID-19 risk factor, in the COV-CLEAR dataset. (E) Distribution of symptomatic subjects in the COV-CLEAR dataset. In each plot, the distributions are grouped with respect to the result of the LFA test: $T = 0$ for a negative test, and $T = 1$ for a positive test.

Appendix B. Additional Plots on Real Data

At the subject level. Fig. 6 presents four examples, where our algorithm either validates or invalidates the result of the test — thereby allowing for the potential flagging of false negatives or negatives.

The first example (Fig. 6 A) is a user that registers a positive test, together with a significant number of symptoms and risk factors.
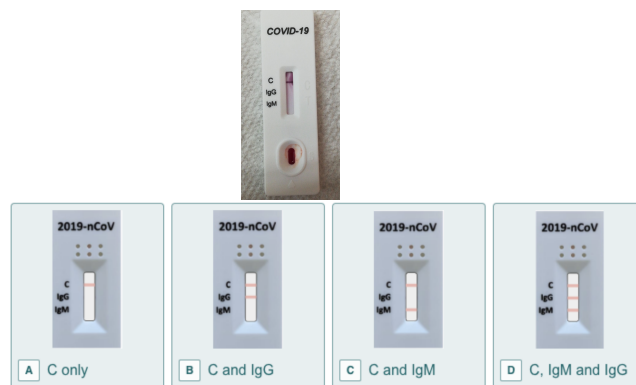
---

4. www.cov-clear.com

Figure 5: Left: Example of test result. Right: Illustrations of a subset of possible test results.

The second user (Fig. 6 B) registers a negative test, while being asymptomatic and with a limited number of risk factors. In both cases, we expect our model to confirm the result of the test, and provide a narrower confidence interval as per the probability of immunity — as confirmed by Fig. 6.

The third and fourth examples showcase instances where the our diagnostic posterior and the test disagree. Subject 108 is a user that registers a negative test, while exhibiting a wide number of known COVID symptoms (dry cough, shortness of breath, fever), but took the test less than 10 days after his illness. Similarly, subject 92 exhibited less symptoms (in particular, no shortness of breath), but lives in a household of three, where all the other members have also fallen sick. While the posteriors reclassify the subjects' diagnosis in each case, the confidence interval associated with the prediction of immunity reflect the uncertainty that is associated with these cases, and flag them as potential false negatives.

At the global level: LFA sensitivity and specificity. The posterior distributions of the model's parameters shed light on the actual accuracy of the LFA tests on our population. Fig. 7 compares the posteriors of the sensitivity and specificity to the priors built from values reported by manufacturers, for:
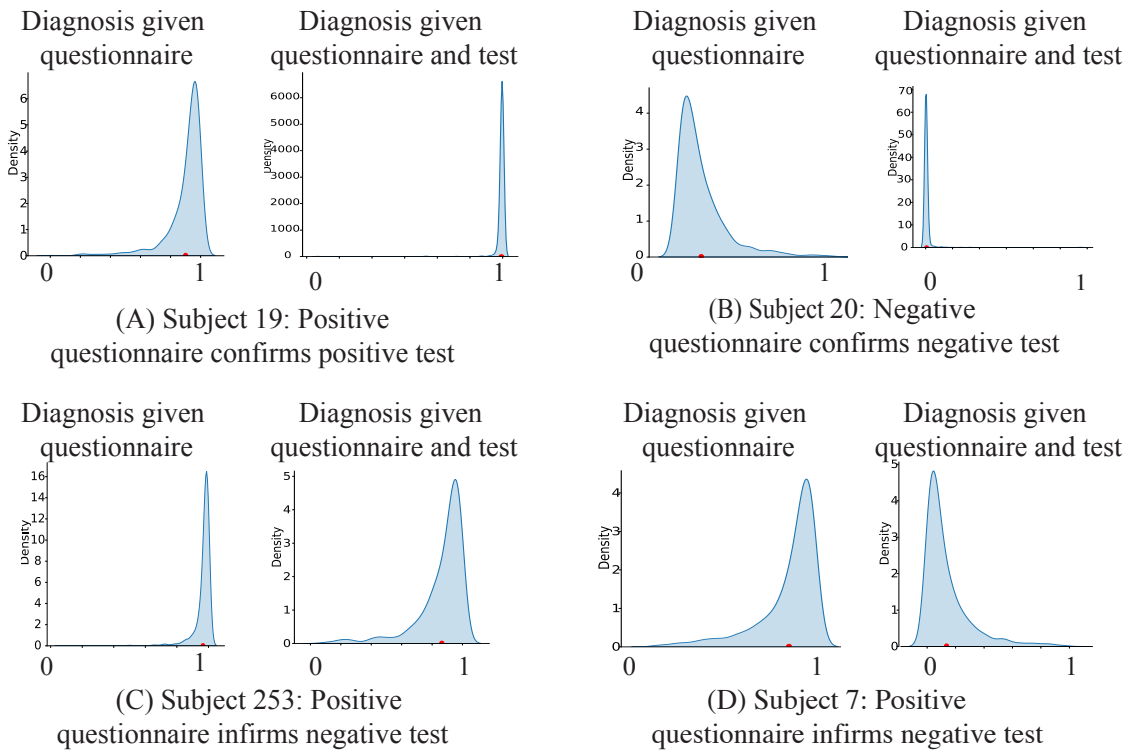
Figure 6: Posterior diagnosis distributions on four selected subjects. In each panel (A-D): the left plot represents the posterior of the diagnosis, given the symptoms and risk factors data reported in the questionnaire, while the right plot is the posterior of the diagnosis, given the symptoms, risk factors, and reported result of the diagnostic test. The red dot represents the expectation of the corresponding distribution.
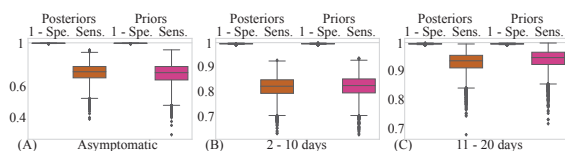
Figure 7: Posteriors of sensitivity and specificity compared to their priors, for three regimes: (A) Asymptomatic, (B) LFA test taken 2-10 days after first symptoms, (C) LFA test taken 11-20 days after first symptoms.

(A) asymptomatic subjects, (B) symptomatic subjects who took the test between 2 to 10 days after their first symptoms, and (C) symptomatic subjects who took the test between 11 to 20 days after their first symptoms. We observe that the StEM has updated the prior distributions of sensitivity and specificity, the most significant update being observed for the asymptomatic subjects.

At the global level: COVID-19 symptoms in the population of interest. Furthermore, our results provide information regarding the most prevalent symptoms for COVID-19. The posterior probability of exhibiting specific symptoms, among symptomatic subjects with positive or negative predicted diagnostic is shown on Fig. 8. We emphasize that these probability distributions are relevant to our population of $n = 117$ healthcare workers, and may vary for studies considering another population.

Appendix C. Application: a COVID-19 immunity study

Preprocessing the data. The prevalence estimates were obtained as the seven-day rolling average of the participants countries, as provided in the Facebook CMU symptoms dataset[5]. **Non binary symptoms.** The data from the COV-CLEAR also has — on
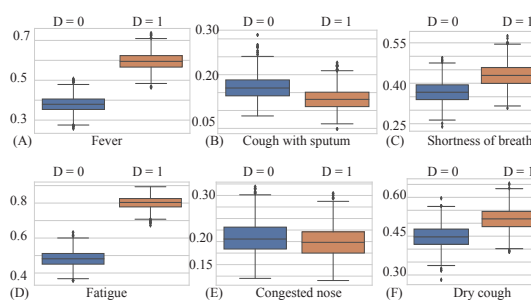

Figure 8: Posterior distributions of the probability of exhibiting selected symptoms, for a symptomatic subject. The probability of exhibiting each symptom (A-F) is plotted for symptomatic subjects with estimated negative ($D = 0$) or positive ($D = 1$) diagnosis.

top of the binary symptoms for which we provided the updates in the main paper — severity index scales — which we choose to model using dirichlet-multinomial conjugate pairs.

Towards extending the results of this study. The population that we are studying in the COV-CLEAR dataset is subject to selection bias: these consist in a population of healthcare workers, and the rate of symptomatic vs asymptomatic people might not be reflective of the one in the general population (there might be more asymptomatic people if this population is younger, etc.). In order to provide potentially more reflective probability of immunity, we make use of the Facebook-CMU dataset, which provides us with more reflective estimates of the case of influenza ($\{D = 0, S = 1\}$ vs covid $\{D = 1, S = 1\}$).

Appendix D. Stochastic Expectation-Maximization

This section presents the derivation of the Stochastic Expectation-Maximization (StEM) algorithm, specifically the proofs of Propositions 4.1 and 4.2 of the main paper. For

---

5. https://www.symptomchallenge.org/about-the-data

the convenience of the reader, we recall the Bayesian model of interest.

D.1. Bayesian generative model

Denote D the true diagnosis of an individual (healthy/sick), T the outcome of the noisy diagnostic test (positive/negative), S the symptomaticity (symptomatic/asymptomatic), X the symptoms exhibited and Y the subject-specific risks factors. The underlying assumption is that given a true diagnosis D, the symptoms X and the diagnostic test outcome T are independent. In other words, the probability of the diagnostic test being a false negative $\mathbb{P}[T = 0|D = 1]$ is independent of the symptoms of a truly infected individual $\mathbb{P}[X|D = 1]$. Similarly, given a diagnosis D, the test outcome T and the exhibited symptoms X are independent of the risk factors Y. We define:

- $\mathbb{P}(D = 1|Y) = \pi_\beta(Y)$ the probability of contracting the disease given risk factors Y,

- $\mathbb{P}[(T = 1|D = 1) = x$ the sensitivity of the diagnostic test,

- $\mathbb{P}[T = 0|D = 0] = 1 - y$ the specificity of the diagnostic test, *i.e.* $y = \mathbb{P}[T = 1|D = 0]$,

- $\mathbb{P}(S = 1|D = 0) = p_0$ the probability of being symptomatic when whilst not having been infected by that specific disease (the symptoms could be due to another illness for instance),

- $\mathbb{P}(S = 1|D = 1) = p_1$ the probability of having been symptomatic upon infection,

- $\mathbb{P}(X_k = 1|S = 1, D = 0) = s_{0k}$ the probability of exhibiting symptom k when not infected,

- $\mathbb{P}(X_k = 1|S = 1, D = 1) = s_{1k}$ the probability of exhibiting symptom k upon infection.

The uncertainty in the test sensitivity and specificity is expressed by putting a prior on x and y, which will be updated during training:
$$\mathbb{P}[T = 1|D = 1] = x \sim \mathcal{B}(\alpha_x, \beta_x),$$
$$\mathbb{P}[T = 1|D = 0] = y \sim \mathcal{B}(\alpha_y, \beta_y). \quad (2)$$
The uncertainty on the symptomaticity, as well as on the appearance of specific symptoms for infected and healthy individuals is expressed with a prior on s_0 and s_1, and updating it when we aggregate more information:
$$\mathbb{P}[S = 1|D = 0] = p_0 \sim \mathcal{B}(\alpha_{p_0}, \beta_{p_0}),$$
$$\mathbb{P}[S = 1|D = 1] = p_1 \sim \mathcal{B}(\alpha_{p_1}, \beta_{p_1}).$$
$$\mathbb{P}[X = 1|D = 0] = s_0 \sim \mathcal{B}(\alpha_{s_0}, \beta_{s_0}), \quad (3)$$
$$\mathbb{P}[X = 1|D = 1] = s_1 \sim \mathcal{B}(\alpha_{s_1}, \beta_{s_1}).$$
Lastly, we model the probability for an individual to contract the disease depending on their risk factors, by modeling the logodds:
$$\log\left(\frac{\pi(Y)}{1 - \pi(Y)}\right) = Y\beta + \epsilon \quad \text{where: } \epsilon \sim N(0, \sigma^2),$$
$$(4)$$
where the parameters β weight the importance of the different components of the risk factors Y (county data, profession, size of household) in contracting the disease. We express our uncertainty on whether a given factor has an impact by putting a prior on β:
$$\beta \sim N(0, \sigma_\beta^2). \quad (5)$$
In this model:

-
$$\zeta = \Big(\alpha_x, \beta_x, \alpha_y, \beta_y, \alpha_{p_0}, \beta_{p_0}, \alpha_{p_1}, \beta_{p_1},$$
$$\{\alpha_{s_{0k}}, \beta_{s_{0k}}\}_k, \{\alpha_{s_{1k}}, \beta_{s_{1k}}\}_k, \{\sigma_{\beta_m}\}_m\Big)$$
are hyper-parameters, considered fixed and known,

- $\theta = \Big(x, y, p_0, p_1, \{s_{0k}, s_{1k}\}_k, \{\beta_m\}_m\Big)$ are the parameters,

- D is a hidden random variable,

- Y, S, X, T are the observed variables.

For simplicity of notations, we write $O = (S, X, T)$ some of the observed variables. The Bayesian model is represented in plate notations in Figure 9.

Our objective is two-fold. *At the patient level*, we compute the posterior distribution of the true diagnosis D_i, informed by the integration of the observed variables O_i. This distribution gives us an estimate of the true diagnosis, through Maximum a Posteriori, as well as a credible interval that expresses our confidence in this diagnosis. *At the global level*, we learn the posterior distributions of the parameters x, y, i.e. the sensitivity and specificity of the test - to be compared to the values given by the providers. We also learn the posterior distributions of the parameters s_0, s_1, i.e. the probability of symptoms with or without the disease - to be compared to the values given by the medical specialists and the CDC. We also learn the posterior distribution of the parameter β, which weights the impact of each risk factor for contracting the disease.

To fulfil this objective, we perform inference in the Bayesian model described in Figure 9 and in the main paper. Since D are hidden variables, we proceed with the Expectation-Maximization (EM) algorithm, specifically its stochastic version.

D.2. Stochastic Expectation Maximization

We seek the Maximum a Posteriori (MAP) of the parameters θ. Given the parameters' estimates, we can compute the posterior distributions of the diagnosiss D_i's. The stochastic EM algorithm allows computing jointly the posterior distribution of the parameters and the hidden variables D_i's, through an iterative procedure described below.

We want to maximise the posterior distribution of the parameters θ:

$$\mathbb{P}(\theta|O_1, ...O_n, Y_1, ..., Y_n)$$
$$= \frac{\mathbb{P}(O_1, ..., O_n|\theta, Y_1, ..., Y_n) \times \mathbb{P}(\theta)}{\mathbb{P}(O_1, ..., O_n|Y_1, ..., Y_n)} \quad (6)$$
$$\propto \mathbb{P}(\theta) \times \Pi_{i=1}^{n} \mathbb{P}(O_i|\theta, Y_i)$$

which translates into maximizing the expression:

$$\ell = \sum_{i=1}^{n} \log \mathbb{P}(O_i|\theta, Y_i) + \log \mathbb{P}(\theta)$$

$$= \sum_{i=1}^{n} \log \sum_{d_i=0,1} \mathbb{P}(O_i, D_i = d_i|\theta, Y_i) + \log \mathbb{P}(\theta)$$

$$= \sum_{i=1}^{n} \log \sum_{d_i=0,1} \mathbb{P}(O_i, D_i = d_i|\theta, Y_i) \frac{\mathbb{P}(D_i = d_i|O_i, \tilde{\theta}, Y_i)}{\mathbb{P}(D_i = d_i|O_i, \tilde{\theta}, Y_i)}$$
$$+ \log \mathbb{P}(\theta)$$

$$\geq \sum_{i=1}^{n} \sum_{d_i=0,1} \mathbb{P}(D_i = d_i|O_i, \tilde{\theta}, Y_i) \log \frac{\mathbb{P}(O_i, D_i = d_i|\theta, Y_i)}{\mathbb{P}(D_i = d_i|O_i, \tilde{\theta}, Y_i)}$$
$$+ \log \mathbb{P}(\theta)$$

$$= \sum_{i=1}^{n} \mathbb{E}_{D_i|O_i, \tilde{\theta}, Y_i} \left[\log \frac{\mathbb{P}(O_i, D_i = d_i|\theta, Y_i)}{\mathbb{P}(D_i = d_i|O_i, \tilde{\theta}, Y_i)} \right] + \log \mathbb{P}(\theta)$$
$$(7)$$

In the above computations, the lower bound is obtained using Jensen inequality. It represents a tangent lower bound of the posterior distribution: $\theta \to \ell(\theta)$ at the given set of parameters $\tilde{\theta}$. On the last line, the expectation is taken for D_i distributed according to $\mathbb{P}(D_i|O_i, \tilde{\theta}, Y_i)$. Following the literature on the stochastic EM algorithm, we compute this expectation by replacing it with its Monte-Carlo estimate, sampling one $\widehat{D_i}$ according to $\mathbb{P}(D_i|O_i, \tilde{\theta}, Y_i)$. For simplicity of notation, we denote:

$$w_i = \mathbb{P}(D_i = d_i|O_i, \tilde{\theta}, Y_i).$$

The approximate lower bound of the posterior of the parameters writes, after sampling:

$$\ell \geq \sum_{i=1}^{n} \log \frac{\mathbb{P}(O_i, D_i = \widehat{D_i}|\theta, Y_i)}{w_i} + \log \mathbb{P}(\theta),$$
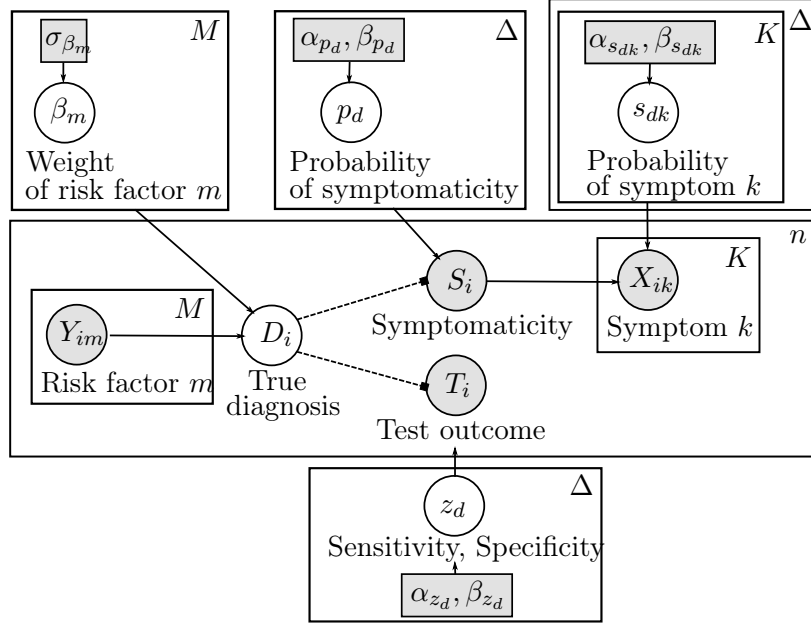$$(8)$$

Figure 9: Hierarchial bayesian network integrating the accuracy uncertainty on the data sources, to estimate the true diagnosis D. The index $k = 1...K$ represents symptom k, while $m = 1...M$ represents risk factor m. The shaded cells represent observed variables or known hyper-parameters. The dashed lines represent switches.

and we only need to maximize the following function $\mathcal{M}$ in its parameters θ:

$$\theta \rightarrow \mathcal{M}(\theta) = \sum_{i=1}^{n} \log \mathbb{P}(O_i, D_i = \widehat{D_i} | \theta, Y_i)$$
$$+ \log \mathbb{P}(\theta).$$

Since $D_i = \widehat{D_i}$ is now fixed, we further decompose the right-hand side of the above inequality.

$$\mathcal{M}(\theta) = \sum_{i=1}^{n} \log \mathbb{P}(T_i, S_i, X_i, D_i = \widehat{D_i} | \theta, Y_i) + \log \mathbb{P}(\theta)$$
$$= \sum_{i=1}^{n} \log \mathbb{P}(T_i, S_i, X_i | D_i = \widehat{D_i}, \theta, Y_i)$$
$$+ \sum_{i=1}^{n} \log \mathbb{P}(D_i = \widehat{D_i} | \theta, Y_i) + \log \mathbb{P}(\theta)$$
$$= \sum_{i=1}^{n} \log \mathbb{P}(T_i | D_i = \widehat{D_i}, \theta, Y_i)$$
$$+ \log \mathbb{P}(S_i, X_i | D_i = \widehat{D_i}, \theta, Y_i)$$
$$+ \sum_{i=1}^{n} \log \mathbb{P}(D_i = \widehat{D_i} | \theta, Y_i) + \log \mathbb{P}(\theta)$$
$$= \sum_{i=1}^{n} \log \mathbb{P}(T_i | D_i = \widehat{D_i}, \theta, Y_i)$$
$$+ \log \mathbb{P}(S_i, X_i | D_i = \widehat{D_i}, \theta, Y_i)$$
$$+ \sum_{i=1}^{n} \log \mathbb{P}(D_i = \widehat{D_i} | \theta, Y_i) + \log \mathbb{P}(\theta)$$
$$= \sum_{i=1}^{n} \log \mathbb{P}(T_i | D_i = \widehat{D_i}, \theta, Y_i)$$
$$+ \log \mathbb{P}(X_i | S_i, D_i = \widehat{D_i}, \theta, Y_i)$$
$$+ \log \mathbb{P}(S_i | D_i = \widehat{D_i}, \theta, Y_i)$$

using the conditional independence of (S_i, X_i) and T_i given D_i. Using the dependency structure of the variables relying on the Bayesian network from Figure 9, we get:

$$\mathcal{M}(\theta) = \sum_{i=1}^{n} \log \mathbb{P}(T_i | D_i = \widehat{D_i}, x, y)$$
$$+ \sum_{i=1}^{n} \log \mathbb{P}(X_i | S_i, D_i = \widehat{D_i}, s_0, s_1)$$
$$+ \log \mathbb{P}(S_i | D_i = \widehat{D_i}, p_0, p_1)$$
$$+ \sum_{i=1}^{n} \log \mathbb{P}(D_i = \widehat{D_i} | \beta, Y_i)$$
$$+ \log \mathbb{P}(x, y) + \log \mathbb{P}(p_0, p_1)$$
$$+ \log \mathbb{P}(s_0, s_1) + \log \mathbb{P}(\beta) \tag{10}$$

As a result, we find the maximum a posteriori of x, y, s_0, s_1 and β separately, by maximizing respectively:

$$M(x, y) = \sum_{i=1}^{n} \log \mathbb{P}(T_i | D_i = \widehat{D_i}, x, y)$$
$$+ \log \mathbb{P}(x, y),$$
$$M(p_0, p_1) = \sum_{i=1}^{n} \log \mathbb{P}(S_i | D_i = \widehat{D_i}, p_0, p_1)$$
$$+ \log \mathbb{P}(p_0, p_1)$$
$$M(s_0, s_1) = \sum_{i=1}^{n} \log \mathbb{P}(X_i | S_i, D_i = \widehat{D_i}, s_0, s_1)$$
$$+ \log \mathbb{P}(s_0, s_1),$$
$$M(\beta) = \sum_{i=1}^{n} \log \mathbb{P}(D_i = \widehat{D_i} | \beta, Y_i) + \log \mathbb{P}(\beta). \tag{11}$$

To summarize, at iteration $(j + 1)$ of the stochastic EM algorithm:

- Stochastic E-step: For each patient i:
  - Compute the posterior distribution of their diagnosis $\mathbb{P}(D_i | O_i, \tilde{\theta}, Y_i)$,
  - Sample $\widehat{D_i}$ from the posterior.

- M-step: Maximize the approximate lower-bound of the parameters' posteri-

ors in $\theta = (x, y, p_0, p_1, s_0, s_1, \beta)$ by maximizing $M(x, y), M(p_0, p_1), M(s_0, s_1)$ and $M(\beta)$. This updates the parameters to:

$$\theta^{(j+1)} = (x^{(j+1)}, y^{(j+1)}, p_0^{(j+1)}, p_1^{(j+1)},$$
$$s_0^{(j+1)}, s_1^{(j+1)}, \beta^{(j+1)}).$$

until convergence in $\theta = (x, y, p_0, p_1, s_0, s_1, \beta)$. At each iteration, this algorithm maximises a (stochastic approximation) of a tangent lower-bound of the posterior distribution of the parameters. Therefore, each iteration increases the posterior distribution of the parameters.

D.3. Auxiliary computation: joint log-likelihood

As an auxiliary computation, we provide the formula for the joint log-likelihood under our model, which we will use in the E- and M-steps.

The joint log-likelihood writes:
$$\mathbb{P}(O, D, \theta | Y)$$
$$= \mathbb{P}(T, S, X, D, x, y, p_0, p_1, s_0, s_1, \beta | Y)$$
$$= \mathbb{P}(T, S, X, x, y, p_0, p_1, s_0, s_1 | D, \beta, Y) \times \mathbb{P}(D, \beta | Y)$$
$$= \mathbb{P}(T, S, X, x, y, p_0, p_1, s_0, s_1 | D)$$
$$\times \mathbb{P}(D | Y, \beta) \times \mathbb{P}(\beta | Y) \tag{12}$$

Using the conditional independence of (S, X) and T given D, we write:
$$\mathbb{P}(O, D, \theta | Y)$$
$$= \mathbb{P}(T, x, y | D) \times \mathbb{P}(S, X, p_0, p_1, s_0, s_1 | D)$$
$$\times \mathbb{P}(D | Y, \beta) \times \mathbb{P}(\beta | Y)$$
$$= \mathbb{P}(T, x, y | D) \times \mathbb{P}(X, s_0, s_1 | S, p_0, p_1, D)$$
$$\times \mathbb{P}(S, p_0, p_1 | D) \times \mathbb{P}(D | Y, \beta) \times \mathbb{P}(\beta | Y)$$
$$= \mathbb{P}(T, x, y | D) \times \mathbb{P}(X, s_0, s_1 | S, D)$$
$$\times \mathbb{P}(S, p_0, p_1 | D) \times \mathbb{P}(D | Y, \beta) \times \mathbb{P}(\beta | Y) \tag{13}$$

D.3.1. AUXILIARY COMPUTATIONS USING THE GENERATIVE MODEL

We separately compute the probabilities in the above formula, using the probability dis-

tributions provided by the generative model. We get, for the term involving the immunoassay test T:

$$\mathbb{P}(T, x, y|D)$$
$$= \mathbb{P}(T|D, x, y) \times \mathbb{P}(x, y|D)$$
$$= \mathbb{P}(T|D, x, y) \times \mathbb{P}(x) \times \mathbb{P}(y)$$
$$= x^{TD}(1-x)^{(1-T)D}y^{T(1-D)}(1-y)^{(1-T)(1-D)}$$
$$\times B(\alpha_x, \beta_x)^{-1}x^{\alpha_x-1}(1-x)^{\beta_x-1}$$
$$\times B(\alpha_y, \beta_y)^{-1}y^{\alpha_y-1}(1-y)^{\beta_y-1}.$$
$$(14)$$

For the term involving the symptoms X, we get:

$$\mathbb{P}(X, s_0, s_1|S, D)$$
$$= \mathbb{P}(X|S, D, s_0, s_1) \times \mathbb{P}(s_0, s_1|D)$$
$$= \mathbb{P}(X|S, D, s_0, s_1) \times \mathbb{P}(s_0) \times \mathbb{P}(s_1)$$
$$= p_0^{(1-S)X}p_1^{(1-S)(1-X)}$$
$$\times s_0^{S(1-D)X}(1-s_0)^{S(1-D)(1-X)}$$
$$\times s_1^{SDX}(1-s_1)^{SD(1-X)}$$
$$\times B(\alpha_{s_0}, \beta_{s_0})^{-1}s_0^{\alpha_{s_0}-1}(1-s_0)^{\beta_{s_0}-1}$$
$$\times B(\alpha_{s_1}, \beta_{s_1})^{-1}s_1^{\alpha_{s_1}-1}(1-s_1)^{\beta_{s_1}-1}$$
$$= \delta_{S=0, X=0}$$
$$+ \delta_{S=1} \times s_0^{(1-D)X}(1-s_0)^{(1-D)(1-X)}$$
$$\times s_1^{DX}(1-s_1)^{D(1-X)}$$
$$\times B(\alpha_{s_0}, \beta_{s_0})^{-1}s_0^{\alpha_{s_0}-1}(1-s_0)^{\beta_{s_0}-1}$$
$$\times B(\alpha_{s_1}, \beta_{s_1})^{-1}s_1^{\alpha_{s_1}-1}(1-s_1)^{\beta_{s_1}-1}$$
$$(15)$$

For the term involving the symptomatic variable S, we get:

$$A = \mathbb{P}(S, p_0, p_1|D)$$
$$= \mathbb{P}(S|p_0, p_1, D) \times \mathbb{P}(p_0) \times \mathbb{P}(p_1)$$
$$= p_0^{(1-D)S}(1-p_0)^{(1-D)(1-S)}$$
$$\times p_1^{DS}(1-p_1)^{D(1-S)}$$
$$\times B(\alpha_{p_0}, \beta_{p_0})^{-1}p_0^{\alpha_{p_0}-1}(1-p_0)^{\beta_{p_0}-1}$$
$$\times B(\alpha_{p_1}, \beta_{p_1})^{-1}p_1^{\alpha_{p_1}-1}(1-p_1)^{\beta_{p_1}-1}.$$
$$(16)$$

For the terms involving the diagnosis D and the parameter β of the logistic regression, we get:

$$\mathbb{P}(D|Y, \beta) = \pi(Y, \beta)^D(1-\pi(Y, \beta))^{1-D}$$
$$\mathbb{P}(\beta) = n(\beta; \sigma_\beta^2),$$
$$(17)$$

where we use the notations:

- $\pi(Y, \beta) = g(Y\beta)$ with g the sigmoid function from the logistic regression,

- $n(\beta; \sigma_\beta^2)$ the probability density function of the Gaussian $\mathcal{N}(0, \sigma_\beta^2)$.

D.3.2. FINAL FORMULA FOR THE JOINT LOG-LIKELIHOOD

We plug the expressions of the probabilities in the joint log-likelihood, and get:

$$\mathbb{P}(O, D, \theta|Y)$$
$$= \mathbb{P}(T, x, y|D) \times \mathbb{P}(X, s_0, s_1|S, D) \times \mathbb{P}(S, p_0, p_1|D)$$
$$\times \times \mathbb{P}(D|Y, \beta) \times \mathbb{P}(\beta|Y)$$
$$\propto x^{TD}(1-x)^{(1-T)D}y^{T(1-D)}(1-y)^{(1-T)(1-D)}$$
$$\times x^{\alpha_x-1}(1-x)^{\beta_x-1}y^{\alpha_y-1}(1-y)^{\beta_y-1}$$
$$\times p_0^{(1-S)X}p_1^{(1-S)(1-X)}s_0^{S(1-D)X}(1-s_0)^{S(1-D)(1-X)}$$
$$\times s_1^{SDX}(1-s_1)^{SD(1-X)}$$
$$\times s_0^{\alpha_{s_0}-1}(1-s_0)^{\beta_{s_0}-1}s_1^{\alpha_{s_1}-1}(1-s_1)^{\beta_{s_1}-1}$$
$$\times p_0^{(1-D)S}(1-p_0)^{(1-D)(1-S)} \times p_1^{DS}(1-p_1)^{D(1-S)}$$
$$\times p_0^{\alpha_{p_0}-1}(1-p_0)^{\beta_{p_0}-1}p_1^{\alpha_{p_1}-1}(1-p_1)^{\beta_{p_1}-1}$$
$$\times \pi(Y, \beta)^D(1-\pi(Y, \beta))^{1-D} \times n(\beta; \sigma_\beta^2),$$
$$(18)$$

where we omit the normalizing constants from the beta distributions.

D.4. Stochastic E-step: Compute posterior of the hidden variables D_i

In the stochastic E-step, we aim to sample $\widehat{D_i}$ from the posterior distribution of the hidden variable D_i. Given the current estimate $\theta^{(j)}$ of the parameters, we compute the posterior of the diagnosis D_i for each patient i:

$$\mathbb{P}(D_i|O_i, \theta^{(j)}, Y_i) \propto \mathbb{P}(D_i, O_i, \theta^{(j)}|Y_i), \quad (19)$$

where we plug the expression of the joint log-likelihood from Equation 44.

Omitting the coefficients that are shared in both formulae for the probability below, we have:

$$\mathbb{P}(D_i = 1|O_i, \theta^{(j)}, Y_i)$$
$$\propto x^{(j)T_i}(1 - x^{(j)})^{(1-T_i)}$$
$$\times s_1^{(j)S_i X_i}(1 - s_1^{(j)})^{S_i(1-X_i)} \times \pi(Y_i, \beta^{(j)})$$
$$\times \times p_1^{S_i}(1 - p_1)^{(1-S_i)}$$
$$\mathbb{P}(D_i = 0|O_i, \theta^{(j)}, Y_i)$$
$$\propto y^{(j)T_i}(1 - y^{(j)})^{(1-T_i)}$$
$$\times s_0^{(j)S_i X_i}(1 - s_0^{(j)})^{S_i(1-X_i)}$$
$$\times (1 - \pi(Y_i, \beta^{(j)})) \times p_0^S(1 - p_0)^{(1-S_i)}, \quad (20)$$

by identification in the expression of the joint log-likelihood from Equation 44.

This shows the proposition:

Proposition 3 *The odds of the posterior of the hidden variable D_i at iteration $(j + 1)$ writes:*
$$\frac{\mathbb{P}(D_i = 1|T_i, S_i, X_i, Y_i, \theta^{(j)})}{\mathbb{P}(D_i = 0|T_i, S_i, X_i, Y_i, \theta^{(j)})}$$
$$= \frac{x^{(j)T_i}(1 - x^{(j)})^{(1-T_i)}}{y^{(j)T_i}(1 - y^{(j)})^{(1-T_i)}}.$$
$$\times \frac{s_1^{(j)S_i X_i}(1 - s_1^{(j)})^{S_i(1-X_i)}}{\times s_0^{(j)S_i X_i}(1 - s_0^{(j)})^{S_i(1-X_i)} \times}$$
$$\times \frac{\pi(Y_i, \beta^{(j)}) \times p_1^{S_i}(1 - p_1)^{(1-S_i)}}{(1 - \pi(Y_i, \beta^{(j)})) \times p_0^{S_i}(1 - p_0)^{(1-S_i)}} \quad (21)$$

Following the methodology of the stochastic EM algorithm, we sample from this posterior to obtain $\widehat{D}_i$.

D.5. M-step: Update model's parameters

In the M-step, we update the parameters of the generative model by maximising the lower-bound of their posterior distribution. We up-date each set of parameters separately, using the expressions in Equation 42.

D.5.1. UPDATE SENSITIVITY AND SPECIFICITY

Given $\widehat{D}_i$'s, we update x, y by maximising:
$$M(x, y) = \sum_{i=1}^{n} \log \mathbb{P}(T_i|D_i = \widehat{D}_i, x, y) + \log \mathbb{P}(x, y). \quad (22)$$

We recognise in this expression the logarithm of the probability distribution of a product of beta distributions in x and in y, omitting the normalisation constants:
$$\Pi_{i=1}^{n}\mathbb{P}(T_i|D_i = \widehat{D}_i, x, y)$$
$$= \Pi_{i=1}^{n}\left(x^{T_i\widehat{D}_i}(1 - x)^{(1-T_i)\widehat{D}_i} \times y^{T_i(1-\widehat{D}_i)}(1 - y)^{(1-T_i)(1-\widehat{D}_i)}\right)$$
$$\mathbb{P}(x, y)$$
$$= x^{\alpha_x - 1}(1 - x)^{\beta_x - 1} \times y^{\alpha_y - 1}(1 - y)^{\beta_y - 1}, \quad (23)$$

such that $M(x, y)$ writes:
$$M(x, y) = \log\left[\mathcal{B}_x\left(\sum_{i=1}^{n} T_i\widehat{D}_i + \alpha_x, \sum_{i=1}^{n}(1 - T_i)\widehat{D}_i + \beta_x\right)\right]$$
$$+ \log\left[\mathcal{B}_y\left(\sum_{i=1}^{n} T_i(1 - \widehat{D}_i)\right.\right.$$
$$\left.\left. + \alpha_y, \sum_{i=1}^{n}(1 - T_i)(1 - \widehat{D}_i) + \beta_y\right)\right]. \quad (24)$$

The updated sensitivity and specificity, $x^{(j+1)}, y^{(j+1)}$, maximize $M(x, y)$. They are the modes of the beta distributions:
$$x^{(j+1)} = \underset{x}{\operatorname{argmax}} \mathcal{B}_x\left(\sum_{i=1}^{n} T_i\widehat{D}_i + \alpha_x, \sum_{i=1}^{n}(1 - T_i)\widehat{D}_i + \beta_x\right)$$
$$y^{(j+1)} = \underset{y}{\operatorname{argmax}} \mathcal{B}_y\left(\sum_{i=1}^{n} T_i(1 - \widehat{D}_i) + \alpha_y,\right.$$
$$\left.\sum_{i=1}^{n}(1 - T_i)(1 - \widehat{D}_i) + \beta_y\right). \quad (25)$$

If $\alpha_x, \beta_x > 1$ and $\alpha_y, \beta_y > 1$, the expression for the modes are:

$$
\begin{aligned}
x^{(j+1)} &= \frac{\sum_{i=1}^n T_i \widehat{D_i} + \alpha_x - 1}{\sum_{i=1}^n T_i \widehat{D_i} + \alpha_x + \sum_{i=1}^n (1 - T_i)\widehat{D_i} + \beta_x - 2} \\
&= \frac{\sum_{i=1}^n T_i \widehat{D_i} + \alpha_x - 1}{\sum_{i=1}^n \widehat{D_i} + (\alpha_x + \beta_x) - 2} \\
y^{(j+1)} &= \frac{\sum_{i=1}^n T_i(1 - \widehat{D_i}) + \alpha_y - 1}{\sum_{i=1}^n T_i(1 - \widehat{D_i}) + \alpha_y + \sum_{i=1}^n (1 - T_i)(1 - \widehat{D_i}) + \beta_y - 2} \\
&= \frac{\sum_{i=1}^n T_i(1 - \widehat{D_i}) + \alpha_y - 1}{\sum_{i=1}^n (1 - \widehat{D_i}) + (\alpha_y + \beta_y) - 2}.
\end{aligned}
\tag{26}
$$

We can have the case $\beta_x, \beta_y < 1$ if the sensitivity or the specificity are very close to 1. In this case, we update the parameters using the expectation of the Beta distribution, instead of the mode.

D.5.2. Update Probabilities of being symptomatic

Given the $\widehat{D_i}$'s, we update p_0, p_1 by maximizing:

$$
M(p_0, s_1) = \sum_{i=1}^n \log \mathbb{P}(S_i | D_i = \widehat{D_i}, p_0, p_1)
$$
$$
+ \log \mathbb{P}(p_0, p_1).
$$

We recognize in this expression the logarithm of the probability distribution of a product of beta distributions in p_0 and in p_1, omitting the normalization constants :

$$
\begin{aligned}
A &= \Pi_{i=1}^n \mathbb{P}(S_i | D_i = \widehat{D_i}, p_0, p_1) \\
&= \Pi_{i=1}^n \Big(p_0^{(1 - \widehat{D_i})S_i}(1 - p_0)^{(1 - \widehat{D_i})(1 - S_i)} \\
&\quad \times p_1^{\widehat{D_i}S_i}(1 - p_1)^{\widehat{D_i}(1 - S_i)} \Big) \\
\mathbb{P}(p_0, p_1) &= p_0^{\alpha_{p_0} - 1}(1 - p_0)^{\beta_{p_0} - 1} \\
&\quad \times p_1^{\alpha_{p_1} - 1}(1 - p_1)^{\beta_{p_1} - 1}
\end{aligned}
\tag{27}
$$

such that $M(p_0, p_1)$ writes:

$$
\begin{aligned}
M(p_0, p_1) = \log \Big[\mathcal{B}_{p_0} \Big(&\sum_{i=1}^n (1 - \widehat{D_i})S_i + \alpha_{p_0}, \\
&\sum_{i=1}^n (1 - \widehat{D_i})(1 - S_i) + \beta_{p_0} \Big) \Big] \\
+ \log \Big[\mathcal{B}_{p_1} \Big(&\sum_{i=1}^n \widehat{D_i} S_i + \alpha_{p_1}, \\
&\sum_{i=1}^n \widehat{D_i}(1 - S_i) + \beta_{p_1} \Big) \Big].
\end{aligned}
\tag{28}
$$

The updated probabilities of being symptomatic, in absence or presence of the disease, $p_0^{(j+1)}, p_1^{(j+1)}$, maximize $M(p_0, p_1)$. They are the modes of the beta distributions:

$$
\begin{aligned}
p_0^{(j+1)} = \underset{x}{\operatorname{argmax}} \; \mathcal{B}_x \Big(&\sum_{i=1}^n (1 - \widehat{D_i})S_i + \alpha_{p_0}, \\
&\sum_{i=1}^n (1 - \widehat{D_i})(1 - S_i) + \beta_{p_0} \Big) \\
p_1^{(j+1)} = \underset{y}{\operatorname{argmax}} \; \mathcal{B}_y \Big(&\sum_{i=1}^n \widehat{D_i}S_i + \alpha_{p_1}, \\
&\sum_{i=1}^n \widehat{D_i}(1 - S_i) + \beta_{p_1} \Big).
\end{aligned}
\tag{29}
$$

If $\alpha_{p_0}, \beta_{p_0} > 1$ and $\alpha_{p_1}, \beta_{p_1} > 1$, the expression for the modes are:

$$
\begin{aligned}
p_0^{(j+1)} &= \frac{\sum_{i=1}^n (1 - \widehat{D_i})S_i + \alpha_{p_0} - 1}{\sum_{i=1}^n (1 - \widehat{D_i})S_i + \alpha_{p_0} + \sum_{i=1}^n (1 - \widehat{D_i})(1 - S_i) + \beta_{p_0} - 2} \\
&= \frac{\sum_{i=1}^n (1 - \widehat{D_i})S_i + \alpha_{p_0} - 1}{\sum_{i=1}^n (1 - \widehat{D_i}) + (\alpha_{p_0} + \beta_{p_0}) - 2} \\
p_1^{(j+1)} &= \frac{\sum_{i=1}^n \widehat{D_i}S_i + \alpha_{p_1} - 1}{\sum_{i=1}^n \widehat{D_i}S_i + \alpha_{p_1} + \sum_{i=1}^n \widehat{D_i}(1 - S_i) + \beta_{p_1} - 2} \\
&= \frac{\sum_{i=1}^n \widehat{D_i}S_i + \alpha_{p_1} - 1}{\sum_{i=1}^n \widehat{D_i} + (\alpha_{p_1} + \beta_{p_1}) - 2}.
\end{aligned}
\tag{30}
$$

We can have the case $\beta_{p_0}, \beta_{p_1} < 1$ if the sensitivity or the specificity are very close to 1. In this case, we update the parameters using the expectation of the Beta distribution, instead of the mode.

D.5.3. UPDATE SYMPTOMS PROBABILITIES

Given $\widehat{D}_i$'s, we update s_0, s_1 by maximizing:

$$M(s_0, s_1) = \sum_{i=1}^{n} \log \mathbb{P}(X_i | S_i, D_i = \widehat{D}_i, s_0, s_1)$$
$$+ \log \mathbb{P}(s_0, s_1), . \quad (31)$$

We recognize in this expression the logarithm of the probability distribution of a product of beta distributions in s_0 and in s_1, omitting the normalization constants:

$$\Pi_{i=1}^{n} \mathbb{P}(X_i | S_i, D_i = \widehat{D}_i, s_0, s_1)$$
$$= \Pi_{i=1}^{n} \Big(\delta_{S_i=0, X_i=0} + \delta_{S_i=1} \times s_0^{(1-\widehat{D}_i)X_i}$$
$$\times (1-s_0)^{(1-\widehat{D}_i)(1-X_i)} s_1^{\widehat{D}_i X_i} (1-s_1)^{\widehat{D}_i(1-X_i)} \Big)$$
$$= \Pi_{i=1, \text{s.t. } S_i=1}^{n} \Big(s_0^{(1-\widehat{D}_i)X_i} (1-s_0)^{(1-\widehat{D}_i)(1-X_i)}$$
$$\times s_1^{\widehat{D}_i X_i} (1-s_1)^{\widehat{D}_i(1-X_i)} \Big) \quad (32)$$

and:

$$\mathbb{P}(s_0, s_1) = s_0^{\alpha_{s_0}-1} (1-s_0)^{\beta_{s_0}-1}$$
$$\times s_1^{\alpha_{s_1}-1} (1-s_1)^{\beta_{s_1}-1}$$

such that $M(s_0, s_1)$ writes:

$$A = M(s_0, s_1)$$
$$= \log \Big[\mathcal{B}_{s_0} \Big(\sum_{i=1, \text{s.t. } S_i=1}^{n} (1-\widehat{D}_i)X_i + \alpha_{s_0},$$
$$\sum_{i=1, \text{s.t. } S_i=1}^{n} (1-\widehat{D}_i)(1-X_i) + \beta_{s_0} \Big) \Big]$$
$$+ \log \Big[\mathcal{B}_{s_1} \Big(\sum_{i=1, \text{s.t. } S_i=1}^{n} \widehat{D}_i X_i + \alpha_{s_1},$$
$$\sum_{i=1, \text{s.t. } S_i=1}^{n} \widehat{D}_i(1-X_i) + \beta_{s_1} \Big) \Big]. \quad (33)$$

The updated probabilities of exhibiting symptoms, in absence or presence of the disease, $s_0^{(j+1)}, s_1^{(j+1)}$, maximize $M(s_0, s_1)$. They are the modes of the beta distributions:

$$s_0^{(j+1)} = \underset{x}{\operatorname{argmax}} \, \mathcal{B}_x \Big(\sum_{i=1, \text{s.t. } S_i=1}^{n} (1-\widehat{D}_i)X_i + \alpha_{s_0},$$
$$\sum_{i=1, \text{s.t. } S_i=1}^{n} (1-\widehat{D}_i)(1-X_i) + \beta_{s_0} \Big)$$

$$s_1^{(j+1)} = \underset{y}{\operatorname{argmax}} \, \mathcal{B}_y \Big(\sum_{i=1, \text{s.t. } S_i=1}^{n} \widehat{D}_i X_i + \alpha_{s_1},$$
$$\sum_{i=1, \text{s.t. } S_i=1}^{n} \widehat{D}_i(1-X_i) + \beta_{s_1} \Big). \quad (34)$$

If $\alpha_{s_0}, \beta_{s_0} > 1$ and $\alpha_{s_1}, \beta_{s_1} > 1$, the expression for the modes are:

$$s_0^{(j+1)} = \frac{\sum_{i=1, \text{s.t. } S_i=1}^{n}(1-\widehat{D}_i)X_i + \alpha_{s_0} - 1}{\sum_{i=1, \text{s.t. } S_i=1}^{n}(1-\widehat{D}_i)X_i + \alpha_{s_0} + \sum_{i=1, \text{s.t. } S_i=1}^{n}(1-\widehat{D}_i)(1-X_i) + \beta_{s_0} - 2}$$
$$= \frac{\sum_{i=1, \text{s.t. } S_i=1}^{n}(1-\widehat{D}_i)X_i + \alpha_{s_0} - 1}{\sum_{i=1, \text{s.t. } S_i=1}^{n}(1-\widehat{D}_i) + (\alpha_{s_0} + \beta_{s_0}) - 2}$$
$$s_1^{(j+1)} = \frac{\sum_{i=1, \text{s.t. } S_i=1}^{n}\widehat{D}_i X + \alpha_{s_1} - 1}{\sum_{i=1, \text{s.t. } S_i=1}^{n}\widehat{D}_i X_i + \alpha_{s_1} + \sum_{i=1, \text{s.t. } S_i=1}^{n}\widehat{D}_i(1-X_i) + \beta_{s_1} - 2}$$
$$= \frac{\sum_{i=1, \text{s.t. } S_i=1}^{n}\widehat{D}_i X_i + \alpha_{s_1} - 1}{\sum_{i=1, \text{s.t. } S_i=1}^{n}\widehat{D}_i + (\alpha_{s_1} + \beta_{s_1}) - 2}. \quad (35)$$

We can have the case $\beta_{s_0}, \beta_{s_1} < 1$ if the probabilities of symptoms are very close to 1. In this case, we update the parameters using the expectation of the Beta distribution, instead of the mode.

D.5.4. UPDATE COEFFICIENTS OF THE RISK FACTORS β

Given the $\widehat{D}_i$'s, we update β by maximizing:

$$M(\beta) = \sum_{i=1}^{n} \log \mathbb{P}(D_i = \widehat{D}_i | \beta, Y_i) + \log \mathbb{P}(\beta)$$
$$= -\sum_{i=1}^{n} \frac{||\widehat{D}_i - g(Y_i\beta)||^2}{2\sigma^2} - \frac{||\beta||^2}{2\sigma_\beta^2} \quad (36)$$

where we omit the normalization constant of the Gaussian distributions. We recognize the loss function of the logistic regression, that we optimize using stochastic gradient descent, with update rule:

$$\beta \leftarrow \beta - \gamma \Big[\sum_i (g(Y_i\beta) - \widehat{D}_i)Y_i + \frac{||\beta||^2}{2\sigma_\beta^2} \Big] \quad (37)$$

This gives the update in β.

D.5.5. SUMMARY OF THE UPDATES

This proposition summarizes the parameters updates.

Proposition 4 *The parameters updates write:*

$$x^{(j+1)} = \frac{\sum_{i=1}^n T_i \widehat{D_i} + \alpha_x - 1}{\sum_{i=1}^n \widehat{D_i} + (\alpha_x + \beta_x) - 2},$$

$$y^{(j+1)} = \frac{\sum_{i=1}^n T_i(1 - \widehat{D_i}) + \alpha_y - 1}{\sum_{i=1}^n (1 - \widehat{D_i}) + (\alpha_y + \beta_y) - 2}$$

$$p_0^{(j+1)} = \frac{\sum_{i=1}^n (1 - \widehat{D_i})S_i + \alpha_{p_0} - 1}{\sum_{i=1}^n (1 - \widehat{D_i}) + (\alpha_{p_0} + \beta_{p_0}) - 2},$$

$$p_1^{(j+1)} = \frac{\sum_{i=1}^n \widehat{D_i}S_i + \alpha_{p_1} - 1}{\sum_{i=1}^n \widehat{D_i} + (\alpha_{p_1} + \beta_{p_1}) - 2}$$

$$s_0^{(j+1)} = \frac{\sum_{i=1,s.t.\ S_i=1}^n (1 - \widehat{D_i})X_i + \alpha_{s_0} - 1}{\sum_{i=1,s.t.\ S_i=1}^n (1 - \widehat{D_i}) + (\alpha_{s_0} + \beta_{s_0}) - 2}$$

$$s_1^{(j+1)} = \frac{\sum_{i=1,s.t.\ S_i=1}^n \widehat{D_i}X_i + \alpha_{s_1} - 1}{\sum_{i=1,s.t.\ S_i=1}^n \widehat{D_i} + (\alpha_{s_1} + \beta_{s_1}) - 2}$$

$$\beta^{(j+1)} = \operatorname{argmin}_\beta \sum_{i=1}^n \frac{||\widehat{D_i} - g(Y_i\beta)||^2}{2\sigma^2} + \frac{||\beta||^2}{2\sigma_\beta^2},$$
(38)

where the minimization on β is performed through stochastic gradient descent.

Appendix E. Stochastic EM with missing T: truncated Bayesian network

We apply the stochastic EM algorithm in the Bayesian model truncated at T. In this model:

-

$$\zeta = \Big(\alpha_{p_0}, \beta_{p_0}, \alpha_{p_1}, \beta_{p_1}, \{\alpha_{s_{0k}}, \beta_{s_{0k}}\}_k,$$
$$\{\alpha_{s_{1k}}, \beta_{s_{1k}}\}_k, \{\sigma_{\beta_m}\}_m\Big)$$

  are hyper-parameters, considered fixed and known,

- $\theta = \Big(p_0, p_1, \{s_{0k}, s_{1k}\}_k, \{\beta_m\}_m\Big)$ are the parameters,

- D is a hidden random variable,

- Y, S, X are the observed variables.

We now writ: $O = (S, X)$.

We still wish to maximize the expression:

$$\ell = \sum_{i=1}^n \mathbb{E}_{D_i|O_i,\tilde{\theta},Y_i}\left[\log \frac{\mathbb{P}(O_i, D_i = d_i|\theta, Y_i)}{\mathbb{P}(D_i = d_i|O_i, \tilde{\theta}, Y_i)}\right]$$
$$+ \log \mathbb{P}(\theta),$$
(39)

under our new notations stated above. The approximate lower bound of the posterior of the parameters still writes, after sampling one $\widehat{D_i}$ according to $\mathbb{P}(D_i|O_i, \tilde{\theta}, Y_i)$:

$$\ell \geq \sum_{i=1}^n \log \frac{\mathbb{P}(O_i, D_i = \widehat{D_i}|\theta, Y_i)}{w_i} + \log \mathbb{P}(\theta),$$
(40)

using the notation: $w_i = \mathbb{P}(D_i|O_i, \tilde{\theta}, Y_i)$. Again, we only need to maximize the following function $\mathcal{M}$ in its parameters θ:

$$\theta \to \mathcal{M}(\theta) = \sum_{i=1}^n \log \mathbb{P}(O_i, D_i = \widehat{D_i}|\theta, Y_i)$$
$$+ \log \mathbb{P}(\theta).$$

Since $D_i = \widehat{D_i}$ is now fixed, we further decompose the right-hand side of the above inequality.

$$\mathcal{M}(\theta) = \sum_{i=1}^n \log \mathbb{P}(S_i, X_i, D_i = \widehat{D_i}|\theta, Y_i) + \log \mathbb{P}(\theta)$$
$$= M(p_0, p_1) + M(s_0, s_1) + M(\beta)$$
(41)

where:

$$M(p_0, p_1) = \sum_{i=1}^n \log \mathbb{P}(S_i|D_i = \widehat{D_i}, p_0, p_1)$$
$$+ \log \mathbb{P}(p_0, p_1)$$

$$M(s_0, s_1) = \sum_{i=1}^n \log \mathbb{P}(X_i|S_i, D_i = \widehat{D_i}, s_0, s_1)$$
$$+ \log \mathbb{P}(s_0, s_1),$$

$$M(\beta) = \sum_{i=1}^n \log \mathbb{P}(D_i = \widehat{D_i}|\beta, Y_i) + \log \mathbb{P}(\beta).$$
(42)

are the same functions as in the case with observed T_i. The only difference is that $\hat{D}_i$ is sampled from $P(D_i|O_i,\tilde{\theta},Y_i)$ where $O_i = (S_i,X_i)$ does not contain T_i.

E.1. Auxiliary computation: log-likelihood in truncated Bayesian model

The joint log-likelihood in the truncated Bayesian model writes:

$$
\begin{aligned}
&\mathbb{P}(O,D,\theta|Y) \\
&= \mathbb{P}(S,X,D,x,y,p_0,p_1,s_0,s_1,\beta|Y) \\
&= \mathbb{P}(S,X,x,y,p_0,p_1,s_0,s_1|D) \\
&\quad \times \mathbb{P}(D|Y,\beta) \times \mathbb{P}(\beta|Y) \\
&= \mathbb{P}(X,s_0,s_1|S,D) \times \mathbb{P}(S,p_0,p_1|D) \\
&\quad \times \mathbb{P}(D|Y,\beta) \times \mathbb{P}(\beta|Y),
\end{aligned}
\tag{43}
$$

which gives the same result as in the non-truncated case, but without the term $\mathbb{P}(T,x,y|D)$. We use the computations from subsection D.3.1 to get the final expression of the loglikelihood: We plug the expressions of the probabilities in the joint log-likelihood, and get:

$$
\begin{aligned}
&\mathbb{P}(O,D,\theta|Y) \\
&\propto p_0^{(1-S)X} p_1^{(1-S)(1-X)} s_0^{S(1-D)X} \\
&\quad \times (1-s_0)^{S(1-D)(1-X)} s_1^{SDX}(1-s_1)^{SD(1-X)} \\
&\quad \times s_0^{\alpha_{s_0}-1}(1-s_0)^{\beta_{s_0}-1} s_1^{\alpha_{s_1}-1}(1-s_1)^{\beta_{s_1}-1} \\
&\quad \times p_0^{(1-D)S}(1-p_0)^{(1-D)(1-S)} \\
&\quad \times p_1^{DS}(1-p_1)^{D(1-S)} \\
&\quad \times p_0^{\alpha_{p_0}-1}(1-p_0)^{\beta_{p_0}-1} p_1^{\alpha_{p_1}-1}(1-p_1)^{\beta_{p_1}-1} \\
&\quad \times \pi(Y,\beta)^D(1-\pi(Y,\beta))^{1-D} \times n(\beta;\sigma_\beta^2),
\end{aligned}
\tag{44}
$$

where we omit the normalizing constants from the beta distributions.

E.2. Stochastic E-step in the truncated model: Compute posterior of the hidden variables D_i

In the stochastic E-step, we aim to sample $\widehat{D}_i$ from the posterior distribution of the hidden variable D_i. Given the current estimate $\theta^{(j)}$ of the parameters, we compute the posterior of the diagnosis D_i for each patient i:

$$
\begin{aligned}
\mathbb{P}(D_i|O_i,\theta^{(j)},Y_i) &= \mathbb{P}(D_i|S_i,X_i,\theta^{(j)},Y_i) \\
&\propto \mathbb{P}(D_i,S_i,X_i,\theta^{(j)}|Y_i),
\end{aligned}
$$

where we plug the expression of the joint log-likelihood of the truncated model.

Omitting the coefficients that are shared in both formulae for the probability below, we have:

$$
\begin{aligned}
&\mathbb{P}(D_i=1|O_i,\theta^{(j)},Y_i) \\
&\quad \propto s_1^{(j)S_iX_i}(1-s_1^{(j)})^{S_i(1-X_i)} \times \pi(Y_i,\beta^{(j)}) \\
&\quad \times p_1^{S_i}(1-p_1)^{(1-S_i)} \\
&\mathbb{P}(D_i=0|O_i,\theta^{(j)},Y_i) \\
&\quad \propto s_0^{(j)S_iX_i}(1-s_0^{(j)})^{S_i(1-X_i)} \\
&\quad \times (1-\pi(Y_i,\beta^{(j)})) \times p_0^S(1-p_0)^{(1-S_i)},
\end{aligned}
\tag{45}
$$

by identification.

Appendix F. Stochastic EM with missing T: Missing and hidden variables

We derive the Stochastic EM algorithm in the full Bayesian model, taking into account the missing variables T_i's (missing for $i = r+1...n$) and hidden variables D_i's. We want to maximize the posterior distribution of the parameters θ:

$$
\begin{aligned}
&\mathbb{P}(\theta|O_1,...O_r,O_{r+1}^M,...,O_n^M,Y_1,...,Y_n) \\
&= \frac{\mathbb{P}(O_1,...,O_r,O_{r+1}^M,...,O_n^M|\theta,Y_1,...,Y_n) \times \mathbb{P}(\theta)}{\mathbb{P}(O_1,...,O_r,O_{r+1}^M,...,O_n^M|Y_1,...,Y_n)} \\
&\propto \Pi_{i=1}^r \mathbb{P}(O_i|\theta,Y_i) \times \Pi_{i=r+1}^n \mathbb{P}(O_i^M|\theta,Y_i) \times \mathbb{P}(\theta)
\end{aligned}
\tag{46}
$$

where we write $O_i = (S_i, X_i, T_i)$ when T_i is available and $O_i^M = (S_i, X_i)$ when T_i is missing.

This translates into maximizing the expression:

$$
\begin{aligned}
\ell = &\sum_{i=1}^{r} \log \mathbb{P}(O_i|\theta, Y_i) + \sum_{i=r+1}^{n} \log \mathbb{P}(O_i^M|\theta, Y_i) \\
&+ \log \mathbb{P}(\theta) \\
\geq &\sum_{i=1}^{r} \mathbb{E}_{D_i|O_i, \tilde{\theta}, Y_i}\left[\log \frac{\mathbb{P}(O_i, D_i = d_i|\theta, Y_i)}{\mathbb{P}(D_i = d_i|O_i, \tilde{\theta}, Y_i)}\right] \\
&+ \sum_{i=r+1}^{n} \mathbb{E}_{D_i, T_i|O_i^M, \tilde{\theta}, Y_i}\left[\log \frac{\mathbb{P}(O_i^M, D_i = d_i, T_i = t_i|\theta, Y_i)}{\mathbb{P}(D_i = d_i, T_i = t_i|O_i^M, \tilde{\theta}, Y_i)}\right] \\
&+ \log \mathbb{P}(\theta),
\end{aligned}
$$

$$(47)$$

where we use Jensen inequality to compute the tangent lower-bounds of $\ell_D = \sum_{i=1}^{r} \log \mathbb{P}(O_i|\theta, Y_i)$ and $\ell_{T,D} = \sum_{i=r+1}^{n} \log \mathbb{P}(O_i^M|\theta, Y_i)$ independently. This is still a valid lower-bound of ℓ at $\tilde{\theta}$ as the sum of the tangent-lower bounds is a tangent-lower bound of the sum.

After sampling, we need to maximize the following function of θ:

$$
\begin{aligned}
\theta \rightarrow \mathcal{M}^{MIS}(\theta) = &\sum_{i=1}^{r} \log \mathbb{P}(O_i, \widehat{D_i} = d_i|\theta, Y_i) \\
&+ \sum_{i=r+1}^{n} \log \mathbb{P}(O_i^M, \widehat{D_i} = d_i, \widehat{T_i} = t_i|\theta, Y_i) \\
&+ \log \mathbb{P}(\theta) \\
= &\mathcal{M}(\theta, n = r) \\
&+ \sum_{i=r+1}^{n} \log \mathbb{P}(O_i^M, \widehat{D_i} = d_i, \widehat{T_i} = t_i|\theta, Y_i)
\end{aligned}
$$

$$(48)$$

where $\mathcal{M}(\theta, n = r)$ is the cost function in the case without missing T.

We compute the second term of the lower-bound, which we denote $\mathcal{M}^{NEW}$:

$$
\begin{aligned}
&\mathcal{M}^{NEW}(\theta) \\
&= \sum_{i=r+1}^{n} \log \mathbb{P}(O_i^M, \widehat{D_i} = d_i, \widehat{T_i} = t_i|\theta, Y_i) \\
&= \sum_{i=r+1}^{n} \log \mathbb{P}(S_i, X_i, \widehat{D_i} = d_i, \widehat{T_i} = t_i|\theta, Y_i) \\
&= \sum_{i=r+1}^{n} \log \mathbb{P}(T_i = \widehat{T_i}|D_i = \widehat{D_i}, \theta, Y_i) \\
&\quad + \log \mathbb{P}(X_i|S_i, D_i = \widehat{D_i}, \theta, Y_i) \\
&\quad + \log \mathbb{P}(S_i|D_i = \widehat{D_i}, \theta, Y_i) \\
&\quad + \sum_{i=1}^{n} \log \mathbb{P}(D_i = \widehat{D_i}|\theta, Y_i)
\end{aligned}
$$

$$(49)$$

which is $\mathcal{M}(\theta, n = (n - r + 1), T_i \rightarrow \widehat{T_i})$. Therefore: $mathcal{M}^{MIS}(\theta)$ is $\mathcal{M}(\subseteq)$ where the missing T_i's have been replaced by theirs imputed value.

We proceed with the Stochastic EM algorithm as follows:

- M-step: Compute the tangent-lower bound, which amounts to:

    - for $i = 1, ..., r$: sample $\widehat{D_i}$ from $\mathbb{P}(D_i|O_i, \tilde{\theta}, Y_i)$,

    - for $i = r + 1, ..., n$: sample $\widehat{D_i}, \widehat{T_i}$ from $\mathbb{P}(D_i, T_i|O_i^M, \tilde{\theta}, Y_i)$,

- E-step: Maximize the lower-bound in θ.

F.1. M-step

For $i = 1, ..., r$, sampling $\widehat{D_i}$ is performed as usual. We focus on sampling $\widehat{D_i}, \widehat{T_i}$ from $\mathbb{P}(D_i, T_i|O_i^M, \tilde{\theta}, Y_i)$.

We compute the posterior of interest:

$$\mathbb{P}(D_i, T_i|O_i^M, \tilde{\theta}, Y_i) \propto \mathbb{P}(D_i, T_i, O_i^M, |\tilde{\theta}, Y_i) \quad (50)$$

We get:
$$\mathbb{P}(D_i = 0, T_i = 0 | O_i^M, \tilde{\theta}, Y_i)$$
$$\propto (1 - y^{(j)}) \times s_0^{(j)S_i X_i}(1 - s_0^{(j)})^{S_i(1-X_i)}$$
$$\times (1 - \pi(Y_i, \beta^{(j)})) \times p_0^S(1 - p_0)^{(1-S_i)}$$
$$\mathbb{P}(D_i = 0, T_i = 1 | O_i^M, \tilde{\theta}, Y_i)$$
$$\propto y^{(j)} \times s_0^{(j)S_i X_i}(1 - s_0^{(j)})^{S_i(1-X_i)}$$
$$\times (1 - \pi(Y_i, \beta^{(j)})) \times p_0^S(1 - p_0)^{(1-S_i)}$$
$$\mathbb{P}(D_i = 1, T_i = 0 | O_i^M, \tilde{\theta}, Y_i)$$
$$\propto (1 - x^{(j)}) \times s_1^{(j)S_i X_i}(1 - s_1^{(j)})^{S_i(1-X_i)}$$
$$\times \pi(Y_i, \beta^{(j)}) \times p_1^{S_i}(1 - p_1)^{(1-S_i)}$$
$$\mathbb{P}(D_i = 1, T_i = 1 | O_i^M, \tilde{\theta}, Y_i)$$
$$\propto x^{(j)} \times s_1^{(j)S_i X_i}(1 - s_1^{(j)})^{S_i(1-X_i)}$$
$$\times \pi(Y_i, \beta^{(j)}) \times p_1^{S_i}(1 - p_1)^{(1-S_i)}.$$
$$(51)$$

These allow to sample $(\widehat{D}_i, \widehat{T}_i)$ according to the posterior, for $i = r + 1, .., n$.

F.2. E-step

The update are the same, except that the missing T_is are replaced by their imputed values.

F.3. Enhancement

In addition, we can consider that we have a different prior on the imputed T_i's. Therefore, we create a new class of sensivitivy/specificity pair, designed for the imputed T_i. The priors are initialized as non-information Beta distribution with parameters $(2, 2)$, and updated at training time.

Appendix G. Additional Plots for Validation on Synthetic Data

This section provides additional details on the validation of the StEM algorithm on synthetic data.

Improvement upon T. Table 1 shows the improvement in the diagnosis' accuracy provided by our method against the sole test T, and highlights the potential strength of harvesting multiple noisy sources of information. The values that we have chosen here for the sensitivity are reflective of the ones that have been reported for the LFA test in the COV-CLEAR study.

Benchmarks. We complete here our discussion of the improvement that our method brings upon Computer-Aided Diagnosis (CAD) standard methods. Fig. 10 compares — for various pairs of sensitivity and specificity. — the diagnosis accuracies of the Stochastic EM algorithm (StEM) and two variants of the EM algorithm: one variant where the parameters' priors are agnostic, or uninformative; and another variant where these priors are computed from the available data, as described in Section 5. This figure demonstrates that learning the parameters' posterior distributions while estimating the diagnosis yields higher prediction accuracy.

Convergence. Fig. 11 shows the distribution of the relative difference between recovered coefficients and ground truth (as a percentage of the ground truth value), showing deviations that are within a few percentages of the true value of the coefficients — thus highlighting the ability of the model to converge to the ground truth parameters.

Fig. 12 shows the average number of convergence steps for different values of sensitivity, specificity and sample sizes. Interestingly, we note that for high values of the sensitivity/specificity, the rate of convergence is much faster. To illustrate the complexity of the algorithm, we also display in Fig. 13 the time required per iteration as a function of the number of samples.

Sensitivity		Specificity			
Real-life	Value	70	80	93	99
Asymptomatic	70	16.2 ±3.9	12.8 ±4.0	8.42 ±3.2	5.64 ±3.2
2-10 days	80	12.3 ±3.2	9.27 ±3.1	4.82±2.9	97.0 ±2.3
11-20 days	93	8.5 ±2.6	6.4 ±2.5	2.3 ±2.1	0.03 ±1.4
21+ days	99.0	8.7 ±2.7	5.7 ±2.0	2.0 ±1.4	0.01 ±0.0

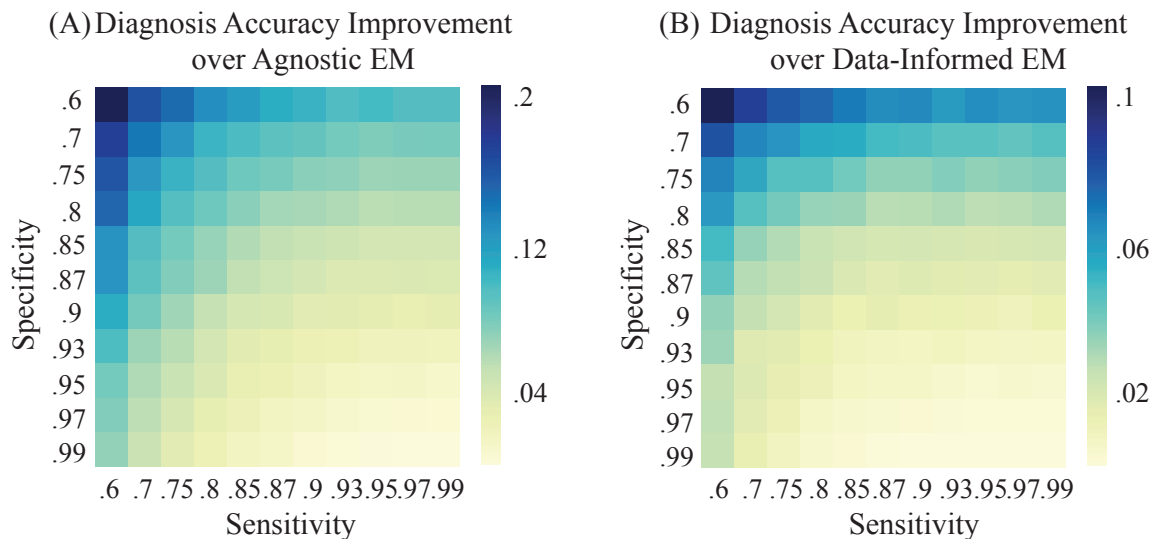Table 1: Gain in accuracy (mean and standard deviation) when using StEM over the sole test

Figure 10: Performance of the StEM algorithm compared to benchmark versions of the EM for n=300 samples, $\sigma = 0.5$, and varying levels of specificity and sensitivity. **(A)** Gain in accuracy with respect to the Data-Agnostic EM described in Section 5. **(B)** Gain in accuracy with respect to the Data-Informed EM described in Section 5.
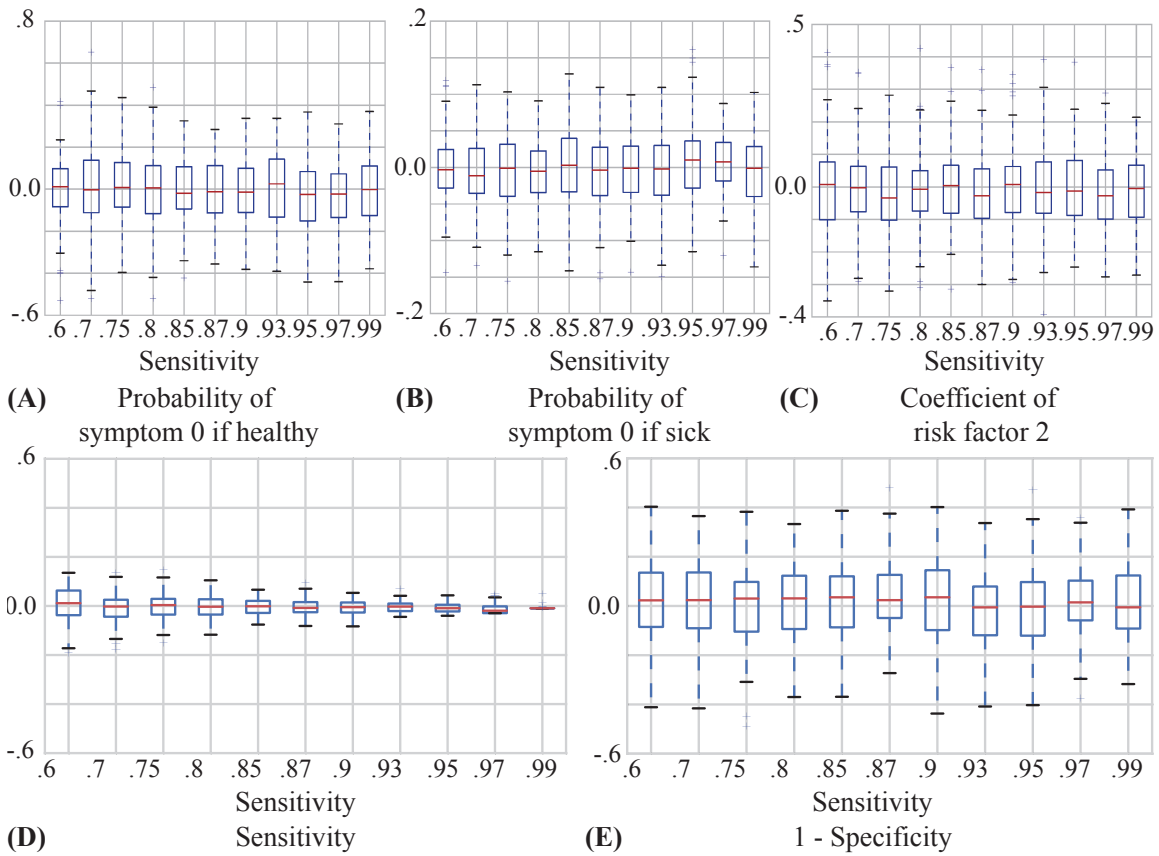
Figure 11: Relative difference between estimated parameters and their ground truth, for $n = 300$, $\sigma = 0.5$ and different values of sensitivity (simulated data).
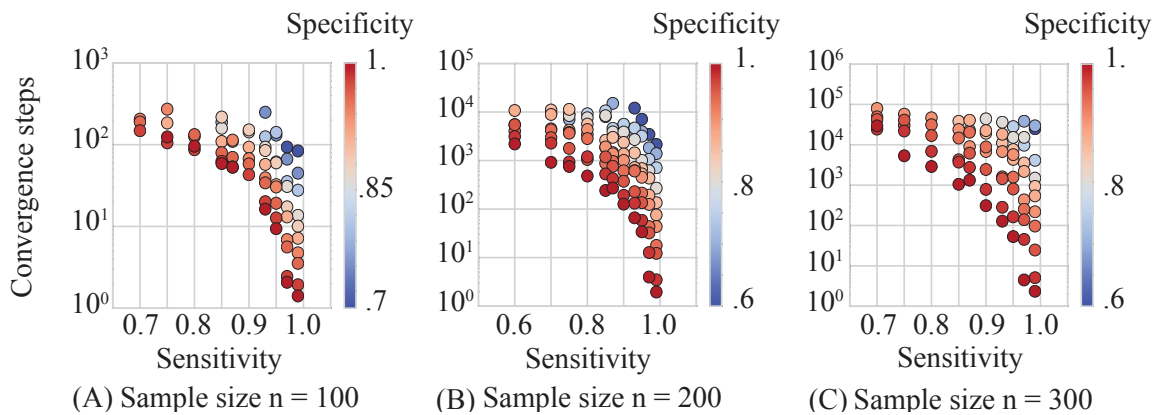

Figure 12: Number of steps until convergence for $\sigma = 0.5$ and different values of sensitivity, specificity and sample sizes (simulated data).
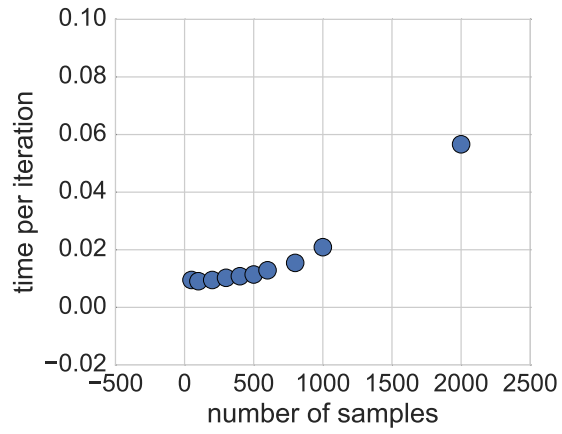
Figure 13: Time (s) per iteration, as the number of samples increases.