Counterfactual diagnosis

Jonathan G. Richens,[1] Ciarán M. Lee,[1,2] and Saurabh Johri[1]

[1] *Babylon Health, London, United Kingdom**

[2] *University College London, United Kingdom*

Causal knowledge is vital for effective reasoning in science and medicine. In medical diagnosis for example, a doctor aims to explain a patient's symptoms by determining the diseases *causing* them. However, all previous approaches to Machine-Learning assisted diagnosis, including Deep Learning and model-based Bayesian approaches, learn by association and do not distinguish correlation from causation. Here, we propose a new diagnostic algorithm based on counterfactual inference which captures the causal aspect of diagnosis overlooked by previous approaches. Using a statistical disease model, which describes the relations between hundreds of diseases, symptoms and risk factors, we compare our counterfactual algorithm to the standard Bayesian diagnostic algorithm, and test these against a cohort of 44 doctors. We use 1763 clinical vignettes created by a separate panel of doctors to benchmark performance. Each vignette provides a non-exhaustive list of symptoms and medical history simulating a single presentation of a disease. The algorithms and doctors are tasked with determining the underlying disease for each vignette from symptom and medical history information alone. While the Bayesian algorithm achieves the accuracy comparable to the average doctor, placing in the top 49% of doctors in our cohort, our counterfactual algorithm places in the top 20% of doctors, achieving expert clinical accuracy. Our results demonstrate the advantage of counterfactual over associative reasoning in a complex real-world task, and show that counterfactual reasoning is a vital missing ingredient for applying machine learning to medical diagnosis.

I. INTRODUCTION

Primary diagnosis, where a doctor chooses between hundreds of diseases to offer an initial diagnosis, is a fundamental problem in healthcare. Errors in primary diagnosis represent a significant burden on global healthcare systems, with at least 5% of patients in the US receiving the wrong diagnosis every year [1]. These errors are particularly common when diagnosing patients with serious medical conditions. An estimated 20% of patients with serious medical conditions are misdiagnosed at the level of primary care [2] and one in three of these misdiagnoses result in serious patient harm [3, 4]. On top of this, over half the global population does not have access to primary healthcare [5].

In recent years, artificial intelligence and machine learning methods have emerged as powerful tools for solving complex problems in diverse domains [6–8]. In particular, Machine Learning assisted diagnosis promises to revolutionise healthcare by providing accurate and accessible diagnoses [9–13]. Diagnostic algorithms can exploit a variety of datasets, from expert opinion and epidemiological data to complex individual risk factors such as genetic information and wearable sensor data [14–18], potentially achieving far higher precision than the cognitive diagnostic models learned by individual doctors and avoiding cognitive biases [19, 20]. These algorithms, which we refer to as *associative* diagnostic algorithms, identify diseases that are strongly correlated with the evidence presented by the patient. For example, Bayesian diagnostic algorithms determine the most likely diagnosis by performing associative Bayesian inference on a statistical disease model (Figure 1). Likewise, deep-learning methods learn to associate patient features with health outcomes but fail to differentiate correlation from causation [21–23]. Despite significant research efforts, diagnostic algorithms are not widely adopted in primary care, largely because they have struggled to achieve the accuracy of human doctors [24–30]. This raises the question—why have diagnostic algorithms failed to live up to their potential?

As noted by Pearl, associative inference is the simplest in a hierarchy of possible inference schemes [23]. Counterfactual inference sits at the top of this hierarchy, and allows one to reason about the consequence of interventions and treatments. Here, we argue that diagnosis is fundamentally a causal inference task. We show that failure to disentangle correlation from causation places strong constraints on the accuracy of associative diagnostic algorithms, which systematically yield spurious diagnoses in certain situations—potentially resulting in sub-optimal care. To resolve this, we present a new causal definition of diagnosis which is closer to the decision making process of clinicians [31], and derive novel *counterfactual* diagnostic algorithms which better captures this causal definition of diagnosis.

Using a statistical disease model, we compare our counterfactual algorithm to the standard associative diagnostic algorithm, and test these against a cohort of 44 doctors. We use 1763 clinical vignettes created by a separate group of doctors to benchmark performance. Each vignette represents a realistic presentation of a patient with a particular disease or condition, containing a non-exhaustive list of evidence including symptoms, medical history, basic demographic information and case notes. A large proportion of the vignettes represent uncommon

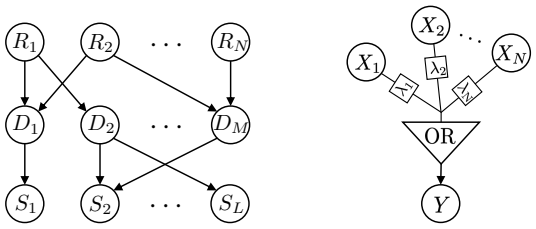
*Electronic address: jonathan.richens@babylonhealth.com

and rare diseases, allowing us to focus more on the cases where diagnostic errors are more common. The algorithms are tasked with determining the model disease in each case from the symptom and past medical history information alone, whilst the doctors receive this evidence along with case notes. On average the cohort of doctors correctly identifies the model disease in 71.98% of vignettes, while the Bayesian algorithm achieves a very similar accuracy of 72.15%, outperforming 51% of doctors. However, our counterfactual algorithm achieves an average score of 76.99%, outperforming 80% of doctors in the study and achieving expert clinical accuracy. We find evidence that the counterfactual algorithm is complimentary to doctors, achieving high accuracy for vignettes that our doctors struggle to diagnose and vice versa. Importantly, the counterfactual algorithm achieves these improvements using the same statistical disease model as the associative algorithm. This backwards compatibility is particularly important as disease models require significant resources to learn. For example, the DMR-DT model was constructed from expert knowledge over three decades [27]. Our algorithms can thus be directly applied to existing diagnostic models, even those outside of medicine [32–35].

II. METHODS

A. Structural causal models for diagnosis

First, we introduce the statistical models used to perform primary diagnosis. These disease models are Bayesian Networks (BNs) that model the relationships between hundreds of diseases, risk factors and symptoms. BNs are widely employed as diagnostic models as they are interpretable [76] and explicitly encode causal relations between variables—a prerequisite for causal and counterfactual analysis [36]. These models typically represent diseases, symptoms and risk-factors as binary nodes that are either on (true) or off (false). We denote true and false with the standard integer notation 1 and 0 respectively. A BN is specified by a directed acyclic graph (DAG) and a joint probability distribution over all nodes which factorises with respect to the graph structure. If there is a directed arrow from node X to Y, then X is said to be a *parent* of Y, and Y to be a *child* of X. A node Z is said to be an *ancestor* of Y is there is a directed path from Z to Y. A simple example BN is shown in Fig. 1 (a), which depicts a BN whose graphical structure describes a three layer network. These models consist of a top layer of risk factor nodes R_i, a middle layer of disease nodes D_j, and a bottom layer of symptoms S_k. As the joint distribution factorises with respect to the graph structure, it is specified by the risk factor priors $p(R_i)$, the disease conditional probability distributions (sometimes referred to as conditional probability tables, or cpts) $p(D_j|R_1,\ldots,R_n)$, and the symptom conditional distributions $p(S_k|D_1,\ldots,D_m)$.



a) 3 layer disease model *b*) Noisy-or cpt

FIG. 1: a) Three layer Bayesian network representing risk factors R_i, diseases D_j and symptoms S_k. b) noisy-OR cpt. Y is the Boolean OR function of its parents, each with an independent probability λ_i of being ignored, removing them from the OR function.

BN disease models have a long history going back to the INTERNIST-1 [25], Quick Medical Reference (QMR) [26, 27], and PATHFINDER [28, 29] systems, with many of the original systems corresponding to noisy-OR networks with only disease and symptom nodes, known as BN2O networks [37]. Recently, three-layer BNs of the form described in the previous paragraph and depicted in Fig. 1 (a) have replaced these two layer models [30]. These models make fewer independence assumptions and allow for disease risk-factors to be included in the diagnostic procedure. Whilst our results will be derived for these models, they can be simply extended to models with more or less complicated dependencies [26, 38].

In the field of causal inference, Bayesian networks are replaced by the more fundamental structural causal models (SCMs), also referred to as functional causal models and structural equation models [39, 40]. SCMs are widely applied and studied, and their relation to other approaches, such as probabilistic graphical models and Bayesian networks, is well understood [36, 41]. The key characteristic of SCMs is that they represent each variable as deterministic functions of their direct causes together with an unobserved, exogenous 'noise' term. That the distribution over the noise term is unknown induces a probability distribution over observed variables. For each variable Y, with parents in the model X, there is a noise term u_Y, with unknown distribution $q(u_Y)$ such that $Y = f(X, u_Y)$ and $p(Y = y|X = x) = \sum_{u_Y : f(X,u_Y)=x} p(U_Y = u_Y)$. By incorporating knowledge of the functional dependencies between variables, SCMs enable us to determine the response of variables to interventions (such as treatments). As we will show in section II E, existing diagnostic BNs such as BN2O networks [37] are naturally represented as SCMs.

B. Posterior ranking

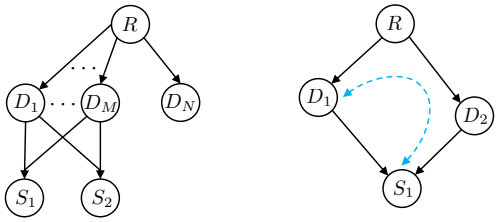
The standard algorithm for performing diagnosis with a disease model, which we refer to as *posterior ranking*, involves ranking candidate diseases by their posterior marginal probabilities. Given a disease model θ

and a patient's set of observed risk-factors $\mathcal{R} = \{R_i\}$ and symptoms $\mathcal{S} = \{S_j\}$, the disease model is used to compute the posterior probability of all model diseases $p(D_k = 1|\mathcal{R}, \mathcal{S}; \theta)$, and the modeled diseases are returned as a ranked list. We assume we are working with a fixed disease model and drop θ from our notation. Although clinical decision making can be informed by other metrics, such as disease severity and harmfulness of treatment [42], disease posteriors are the key ingredient supplied by the disease model to inform the diagnosis.

Whilst the posterior quantifies the likelihood of a given disease, it cannot differentiate causal and acausal correlations. Using the posterior to explain an observation often leads to spurious conclusions in all but the simplest models. To see this, consider the toy disease model shown in Fig 2. a). In the case where we observe $\mathcal{E} = \{S_1 = 1, S_2 = 1\}$, and $p(D_i = 1|\mathcal{E}) \sim p(D_j = 1|\mathcal{E})$ $\forall\ i, j \in 1, \dots M$, then diseases $D_1, \dots, D_M$ will typically have small posteriors for large M. This causes $D_1, \dots, D_M$ to be strongly anti-correlated given $\mathcal{E}$, as these diseases compete to explain the evidence $\mathcal{E}$. This phenomenon is known as 'explaining away' [43]. However, it is very likely that at least one of $D_1, \dots, D_M$ is present as $p(\text{any}\{D_i = 1\}_{i=1}^{M}|\mathcal{E}) \sim 1$. As a consequence the risk factor posterior can be large, $p(R = 1|\mathcal{E}) \sim p(R = 1|\text{any}\{D_i = 1\}_{i=1}^{M}) \sim 1$, as R is capable of explaining all $D_1, \dots, D_M$. If disease D_N has a strong enough association with R, e.g. $p(D_N = 1|R = 1) > p(D_i = 1|R = 1)$, for $i = 1, \dots, M$, then D_N can end up with the largest posterior, $p(D_N = 1|\mathcal{E}) > p(D_i = 1|\mathcal{E})$ for all $i = 1, \dots, M$, despite it being impossible that D_N is causing the observed symptoms.

This is an example of confounding, where the latent variable R causes a spurious association between the symptom evidence and D_N [44]. Note that this spurious association is due to a flawed interpretation of the posterior, rather than a flaw in the model. Furthermore, note that even if R is observed the problem can persist (for example if we observe $R = 1$), potentially resulting in D_N being put forward as a likely diagnosis despite having no symptom evidence. As a real-world example, consider an elderly smoker who reports chest pain, nausea, and fatigue. A good doctor will present a diagnosis that is both likely and relevant given the evidence (such as angina). Although it may also be true that this patient is likely to belong to a population that frequently suffer from other diseases such as emphysema, this disease has little to do with the evidence presented and should not be put forward as a diagnosis. Here, emphysema is a disease whose posterior weight is derived primarily from 'back-door paths' [45], as with D_N in Fig 2 a).

In Fig 2. b) we show another example of confounding. In this case, suppose D_1 has a positive causal association with S_1, i.e. $D_1 = 1$ makes $S_1 = 1$ more likely. In this case, the observation $S_1 = 0$ has a negative causal association with $D_1 = 1$. However, back-door paths between D_1 and S_1 (e.g. latent risk-factors and diseases) can reverse this association in a phenomena known as Simpson's paradox [44]. As a result, observing $S_1 = 0$ can make $D_1 = 1$ more likely. However, causally we know that $D_1 = 1$ is a very poor explanation for the evidence $S_1 = 0$. If the diagnosis is made using the posterior alone, D_1 could be put forward as a diagnosis despite there being negative evidence for this disease being the cause of the patient's symptoms.



a) Non-causal association *b*) Back-door confounding

FIG. 2: a) latent risk-factor confounding results in correlations between latent diseases and symptom evidence where no causal link exists b) Latent diseases and risk factors form back-door paths between diseases and symptoms (blue arrow). This results in Simpsons-paradox type confounding [44]

In general, diseases will have a both causal (front door) and acausal (back-door) connections to the observed symptom evidence. It is not sufficient to ignore diseases like D_N in Fig 2. (a) that are not ancestors of the symptom evidence. We need a formal approach to diagnosis that can resolve these issues by disentangling correlation from causation. Our examples show that ignoring causal information can lead to a spurious diagnosis. In practice, human clinicians will not present completely spurious diseases in a diagnosis, i.e. diseases that cannot generate the patient's symptoms, and we can conclude that human clinicians are making use of this causal information in their diagnosis. We now propose an algorithm that can also exploit this information.

C. Beyond associative diagnosis

In section II B, we showed diagnostic algorithms that rely solely on posterior ranking can yield sub-optimal or spurious diagnoses. To resolve this, we first consider how to best define diagnosis. The Oxford English dictionary defines it as "the identification of the nature of an illness or other problems by examination of the symptoms". That is, given the symptoms and risk factors presented by the patient, a doctor attempts to determine the diseases that are the best explanation—the *most likely underlying cause*—of the symptoms presented. We propose the following causal definition for diagnosis,

The identification of the diseases that are most likely to be causing the patient's symptoms, given their medical history

This definition suggests the following three minimal desiderata that should be satisfied by any diagnostic algo-

rithm aiming to capture the likelihood, $\mathcal{M}(D_k, E)$, that a disease, D_k, is causing a patient's symptoms given evidence, $\mathcal{E}$,

i) The likelihood that a disease D_k is causing a patient's symptoms should be proportional to the posterior likelihood of that disease $\mathcal{M}(D_k, E) \propto p(D_k = 1|\mathcal{E})$ (*consistency*),

ii) Any disease D_k that cannot cause any of the patient's observed symptoms $\mathcal{S}_+$ should not be included in a diagnosis, $D_k \notin \text{Ancestors}(\mathcal{S}_+) \implies \mathcal{M}(D_k, \mathcal{E}) = 0$ (*causality*),

iii) Diseases that explain a greater number of the patient's symptoms should be more likely (*simplicity*).

The justification for these desiderata is as follows. Desiderata i) states that the likelihood that a disease explains the patient's symptoms depends on the likelihood that the patient has the disease in the first place. Desiderata ii) states that if there is no causal mechanism whereby disease D_k could have generated any of the positive symptoms presented (directly or indirectly), then D_k cannot constitute causal explanation of the symptoms and should be disregarded. Desiderata iii) incorporates the principle of Occam's razor—favouring simple diagnoses with few diseases that can explain many of the symptoms presented. It is clear from section II B that posterior ranking only satisfies the first desiderata, violating the last two.

D. Counterfactual diagnosis

To quantify the likelihood that a disease is causing the patient's symptoms, we employ counterfactual inference [46–48]. Counterfactuals can test whether certain outcomes *would have* occurred had some precondition been different. Given evidence $\mathcal{E} = e$ we calculate the likelihood that we would have observed a different outcome $\mathcal{E} = e'$—*counter to the fact* $\mathcal{E} = e$—had some hypothetical intervention taken place. The counterfactual likelihood is written $P(\mathcal{E} = e' \,|\, \mathcal{E} = e, \text{do}(X = x))$ where $\text{do}(X = x)$ denotes the intervention that sets variable X to the value $X = x$, as defined by Pearl's calculus of interventions [36] (see appendix C for formal definitions).

Counterfactuals provide us with the language to quantify how well a disease hypothesis $D_i = 1$ explains symptom evidence $S = 1$ by determining the likelihood that the symptom would not be present if we were to intervene and 'cure' the disease by setting $\text{do}(D_i = 0)$, given by the counterfactual probability $P(S = 0 \,|\, S = 1, \text{do}(D_i = 0))$. If this probability is high, $D_i = 1$ constitutes a good causal explanation of the symptom. Note that this probability refers to two contradictory states of S and so cannot be represented as a standard posterior [36, 39]. In appendix C we describe how these counterfactual probabilities are calculated.

Inspired by this example, we propose two counterfactual diagnostic measures, which we term the *expected disablement* and *expected sufficiency*. We show in Theorem 1 at the end of this section that both measures satisfy all three desiderata from section II C.

Definition 1 (Expected disablement)**.** *The expected disablement of disease D_k determines the number of positive symptoms that we would expect to switch off if we intervened to turn off D_k,*

$$\mathbb{E}_{dis}(D_K, \mathcal{E}) := \sum_{\mathcal{S}'_+, \mathcal{S}'_-} \left|\mathcal{S}_+ \setminus \mathcal{S}'_+\right| p(\mathcal{S}'_+, \mathcal{S}'_- | \mathcal{E}, do(D_k = 0)) \tag{1}$$

where $\mathcal{E}$ is the factual evidence, $\mathcal{S}_+$ and $\mathcal{S}_-$ are the sets of factual positively and negatively evidenced symptoms respectively. $\mathcal{S}'_+$ is the set of counterfactual symptoms that remain positive on following the intervention $do(D_k = 0)$.

The expected disablement derives from the notion of necessary cause [49], whereby D is a necessary cause of S if $S = 1$ if and only if $D = 1$. The expected disablement therefore captures how well disease D_k alone can explain the symptoms, as well as the likelihood that treating D_k alone will alleviate the patient's symptoms.

Definition 2 (expected sufficiency)**.** *The expected sufficiency of disease D_k determines the number of positively evidenced symptoms we would expect to persist if we intervene to switch off all other possible causes of the symptoms,*

$$\mathbb{E}_{suff}(D_K, \mathcal{E}) := \sum_{\mathcal{S}'} \left|\mathcal{S}'_+\right| p(\mathcal{S}' | \mathcal{E}, do(\mathsf{Pa}(\mathcal{S}_+) \setminus D_k = 0)) \tag{2}$$

where $\mathsf{Pa}(\mathcal{S}'_+)$ denotes the set of all parents of the set of counterfactual positively evidenced symptoms $\mathcal{S}'_+$ excluding D_k. The expectation is calculated over all possible counterfactual symptom states $\mathcal{S}'$, where $\mathcal{S}'_+$ denotes the set of positive symptoms in the counterfactual symptom state. $\mathcal{E}$ denotes the set of all factual evidence.

The expected sufficiency derives from the notion of sufficient cause [49], whereby D is a sufficient cause of S if the presence of D implies the subsequent occurrence of S but, as S can have multiple causes, the presence of S does not imply the prior occurrence of D. Typically, diseases are sufficient causes of symptoms, and sufficient causes can be quantified by controlling for all other sufficient causes. In our case, we perform counterfactual interventions to remove all possible causes of the symptoms (both diseases and exogenous influences), and consider all counterfactual symptom states that could have occurred in this scenario. As these counterfactual symptom values have only a single possible cause—the disease D—any symptoms remaining must have been caused by D, allowing us to quantify the number of symptoms that we can expect to have been caused by D.

Theorem 1 (Diagnostic properties of expected disablement and expected sufficiency)**.** *Expected Disablement and expected sufficiency satisfy the three desiderata from section II C*

The proof is provided in appendices E and G.

E. Noisy-OR and twin network diagnostic models

When constructing disease models it is common to make additional modelling assumptions beyond those implied by the DAG structure. The most widely used of these are 'noisy-OR' models [26], as they closely fit our beliefs about how diseases develop [50, 51], and allow for large BNs to be described by a number of parameters that grows linearly with the size of the network [52, 53]. We now derive expressions for the expected disablement and expected sufficiency for these models, which allow these measures to be determined using standard inference techniques.

Under the noisy-OR assumption, a parent X_i activates its child Y (causing $Y = 1$) if i) the parent is on, $X_i = 1$, and ii) the activation does not randomly fail. The probability of failure, conventionally denoted as $\lambda_{X_i,Y}$, is independent from all other model parameters. The 'OR' component of the noisy-OR states that the child is activated if *any* of its parents successfully activate it. Concretely, the values of Y is the Boolean OR function $\vee$ of its parents activation functions, $y = \vee_i f(x_i, u_i)$, where the activation functions take the form $f(x_i, u_i) = x_i \wedge \bar{u}_i$, where $\wedge$ denotes the Boolean AND function, $x_i \in \{0, 1\}$ is the state of a given parent X_i and $u_i \in \{0, 1\}$ is a latent noise variable ($\bar{u}_i := 1 - u_i$) with a probability of failure $p(u_i = 1) = \lambda_{X_i,Y}$. The noisy-OR model is depicted in Fig 1. b). As we show in appendix B, noisy-OR models are naturally formulated in the SCM framework, described in section II A. For further details on noisy-OR disease modelling see appendix B.

To efficiently evaluate our counterfactual diagnosis measures we employ the twin networks method for computing counterfactuals, outlined in [54, 55]. This method represents real and counterfactual variables together in a single SCM—the twin network—from which counterfactual probabilities can be computed using standard inference techniques. This approach greatly amortizes the inference cost of calculating counterfactuals compared to the standard approach of abduction, action and prediction [36], which is intractable for large SCMs. We refer to these diagnostic models as twin diagnostic networks, see appendix C for further details. We now derive expressions for the expected disablement and expected sufficiency for 3-layer noisy-OR disease models in terms of corrections to the standard posterior probabilities.

Theorem 2. *For 3-layer noisy-OR BNs (formally described in appendix B-C), the expected sufficiency and expected disablement of disease D_k are given by*

$$\frac{\sum_{\mathcal{Z}\subseteq\mathcal{S}_+}(-1)^{|\mathcal{Z}|}p(\mathcal{S}_-=0,\mathcal{Z}=0,D_k=1|\mathcal{R})\tau(k,\mathcal{Z})}{p(\mathcal{S}_\pm|\mathcal{R})} \tag{3}$$

where for the expected sufficiency

$$\tau(k,\mathcal{Z}) = \sum_{S\in\mathcal{S}_+\setminus\mathcal{Z}}(1-\lambda_{D_k,S}) \tag{4}$$

and for the expected disablement

$$\tau(k,\mathcal{Z}) = \sum_{S\in\mathcal{Z}}\left(1-\frac{1}{\lambda_{D_k,S}}\right) \tag{5}$$

where $\mathcal{S}_\pm$ denotes the positive and negative symptom evidence, $\mathcal{R}$ denotes the risk-factor evidence, and $\lambda_{D_k,S}$ is the noise parameter for D_k and S.

The proof is provided by theorem 4 in appendix D and by theorem 5 in appendix F. Note that (3) recovers the standard posterior $p(D_k = 1|\mathcal{E})$ in the limit that $\tau(D_k, \mathcal{Z}) \to 1 \; \forall \; \mathcal{Z}$.

III. ACHIEVING EXPERT PERFORMANCE IN PRIMARY DIAGNOSIS

A common approach to validating diagnostic algorithms is to use real clinical cases labeled by their diagnosis [9–13]. A key limitation of this approach is the difficulty in defining the ground truth diagnosis in real cases, where diagnostic errors result in mislabellings. This problem is particularly pronounced for the diagnoses that occur during primary care due to many factors including; the large number of candidate diseases and hence diagnostic labels, incomplete or inaccurate recording of case data in medical health records, high diagnostic uncertainty and ambiguity, and biases such as the training and experience of the clinician who performed the diagnosis. To resolve these issues, a standard method for assessing doctors is through the examination of simulated diagnostic cases or *clinical vignettes* [56]. A clinical vignette simulates a typical diagnostic case for a model disease or diseases, and doctors are assessed on their ability to return an appropriate diagnosis for a given vignette. Clinical vignettes generated by expert panels of clinicians are often more robust to errors and biases than labeled data sets such as medical health records, as the task is generative—simulating a disease given its known properties—rather than discriminative—diagnosing an unknown disease. This approach has been found to be effective for evaluating human doctors [56–59] and comparing the clinical accuracy of doctors to symptom checker algorithms [24, 30, 60, 61].

To evaluate our algorithms, we first construct a test set of 1763 clinical vignettes, generated by a seprate

panel of doctors qualified at least to the level of general practitioner [77]. Each vignette represents a realistic presentation of a patient with a single disease or condition, containing a list of evidence including symptoms, medical history, and basic demographic information such as age and birth gender [30]. The evidence list is non-exhaustive, reflecting the partial information typically available to a doctor during a primary consultation. Where possible, symptoms and risk factors matches those in our statistical disease model (see below) to allow the system to recognise these variables as evidence. However, to avoid biasing our study by only including evidence that is compatible with our disease model, the vignette author is allowed to include any additional clinical information as case notes, which are available to the doctors in our experiments. An example vignette is shown in Figure 4 b). Each vignette is authored by a single doctor and then verified by multiple doctors to ensure that it represents a realistic diagnostic case.

For a given case, the algorithm is provided with the evidence list with the model disease masked. The algorithm returns a diagnosis in the form of a full ranking of all diseases in the disease model, based on the posterior marginal probabilities (for the associative algorithm) or the expected disablement or expected sufficiency (for the counterfactual algorithms). These rankings are then used to compare the diagnostic accuracy of the algorithms.

In all experiments the counterfactual and associative algorithms use identical disease models to ensure that any difference in diagnostic accuracy is solely due to the algorithm used. The disease model used is a three layer noisy-or diagnostic BN as described in sections II B and II E and appendix B. The BN is parameterised by a team of doctors and epidemiologists [30, 61]. The DAG of the disease model, which encodes the causal structure of the diseases, risk factors and symptoms, is determined by a panel of doctors. The prior probabilities of diseases and risk factors are obtained from epidemiological data, where available. Conditional probabilities are obtained through elicitation from multiple independent medical sources and doctors [78]. The expected disablement and expected sufficiency are calculated using Theorem 2.

Our first experiment compares the accuracies of our counterfactual diagnostic algorithms to the associative algorithm (posterior ranking). We diagnose each of the 1763 vignettes using the posterior, expected disablement and expected sufficiency to produce a full ranking of all diseases. The top k accuracy is then calculated as fraction of the 1763 diagnostic vignettes where the model disease is present in the k top ranked diseases returned by the algorithm. The results are presented in figure 3. The expected disablement and expected sufficiency give almost identical rankings on our test set, and for the sake of clarity we present the results for the expected sufficiency alone, which we refer to as the counterfactual algorithm. A complete table of results is present in Appendix H.

For $k = 1$—returning the top ranked disease—the two algorithms perform similarly, with the counterfactual algorithm improving on the associative algorithm's accuracy by 2%. For $k > 1$ the performance of two algorithms diverge, with the counterfactual algorithm giving a large reduction in the error rate over the associative algorithm. For $k = 5$, the counterfactual algorithm reduces the number of misdiagnoses by 30% compared to the associative algorithm. The fact that the improvement increases for large k suggests that the counterfactual algorithm is removing spurious diseases from the ranking.

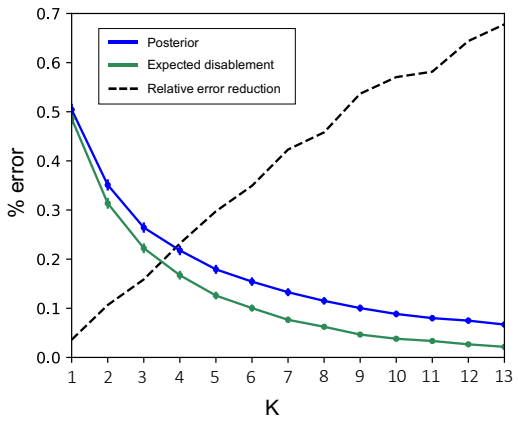


FIG. 3: **Top k accuracy of Bayesian and counterfactual algorithms**. Figure shows the top k error (1 - accuracy) of the counterfactual and associative algorithms over all 1763 vignettes v.s k. The black dashed line shows the relative reduction in error when switching from the associative to counterfactual algorithm, given by $1 - e_c/e_a$ where e_a is the error rate of the associative algorithm, and e_c is the error rate of the counterfactual algorithm.

	vignettes					
	All	VCommon	Common	Uncommon	Rare	VRare
N	1763	139	437	564	375	224
Mean (A)	3.65	2.90	2.92	3.37	4.17	4.94
Mean (C)	2.77	2.35	2.25	2.58	3.18	3.71
Wins (A)	43	6	9	12	14	2
Wins (C)	435	23	85	127	109	82
Draws	1285	110	343	425	252	140

TABLE I: **Mean position of model disease in ranking stratified by rareness of disease**. Table shows the mean position of the model disease for the associative (A) and counterfactual (C). The results for expected disablement as almost identical to the expected sufficiency and are included in the appendices. Results are stratified over the rareness of the disease (given the age and gender of the patient), where VCommon = Very common and VRare = very rare, and All is over all 1763 vignettes regardless of disease rarity. N is the number of vignettes belonging to each rareness category. Wins (X) is the number of vignettes where algorithm X ranked the model disease higher than the is counterpart, and Draws is the number of vignettes where the two algorithms ranked the model disease in the same position. For full results including uncertainties see appendix H.

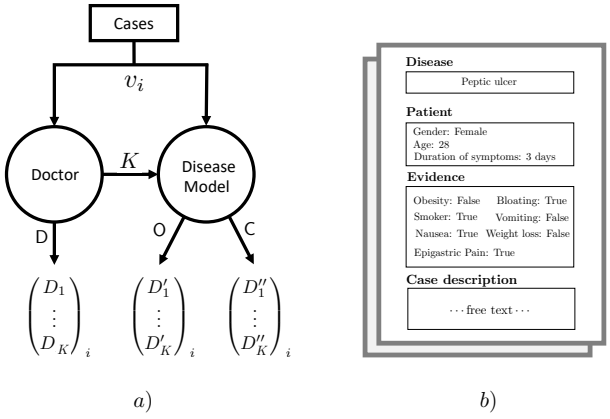


FIG. 4: Figure shows the setup of our experiment. Vignettes v_i are drawn at random and are first passed to the doctor D, who returns a diagnosis of size k where k is chosen by the doctor on a case-by-case basis. The counterfactual C and associative A algorithms are passed the same vignette and perform a complete ranking of all model diseases. The top k of these are then selected, matching the precision of the given doctor for each vignette. The experiment is run independently for each doctor. Figure b) depicts an example of a medical vignette.

A simple method for comparing two rankings is to compare the position of the model disease in the rankings. Across all 1763 vignettes we found that the counterfactual algorithm ranked the model disease higher than the associative algorithm in 24.8% of vignettes, and lower in only 2.4% of vignettes. On average the model disease is ranked in position 2.77±3.12 by the counterfactual algorithm, a substantial improvement over 3.65±4.79 for the associative algorithm (see Table I).

In table I we stratify the vignettes into very common, common, uncommon, rare and very rare depending on the the prior incidence rates of the model disease. While the counterfactual algorithm achieves significant improvements over the associative algorithm for both common and rare diseases, the improvement is particularly large for rare diseases, achieving a higher ranking for 36.7% of these vignettes. This is particularly important as rare diseases are typically harder to diagnose and include many serious conditions where diagnostic errors have the greatest consequences.

The second experiment compares the counterfactual and associative algorithms to a cohort of 44 doctors. Each doctor and is assigned a set of at least 50 clinical vignettes (average 159), and returns an independent diagnosis for each vignette in the form of a partially ranked list of k diseases, where the size of the list k is chosen by the doctor on a case-by-case basis. The average diagnosis size is 2.58 diseases. For a given doctor, and for each case diagnosed by the doctor, the associative and counterfactuals algorithms are supplied with the same evidence (excluding the free text case description) and each returns a top k diagnosis, where k is the size of the diagnosis provided by the doctor for this case. Hence the algorithms shadow each doctor, providing a second-opinion diagnosis with the same number of diseases as the doctor for each case. The experimental setup is depicted in figure 4 a). Matching the precision of the doctor for every case allows us to compare the accuracy of the doctor and the algorithms without constraining the doctors to give a fixed number of diseases in each diagnosis. This is important as doctors will naturally vary the size k of their diagnosis to reflect their uncertainty in the diagnostic case.

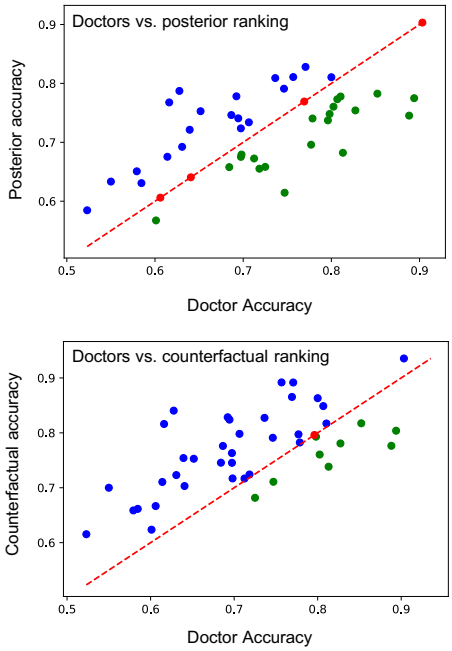


FIG. 5: **Mean accuracy of each doctor compared to Bayesian and counterfactual algorithms.** Figure shows the mean accuracy for each of the 44 doctors, compared to the posterior ranking (top) and expected sufficiency ranking (bottom) algorithms. The line $y = x$ gives a reference for comparing the accuracy of each doctor to the algorithm shadowing them. Points above the line correspond to doctors who achieved a lower accuracy than the algorithm (blue), points on the line are doctors that achieved the same accuracy as the algorithm (red), and below the line are doctors that achieved higher accuracy than the algorithm (green). The linear correlation can be explained by the variation in the difficulty of the sets of vignettes diagnosed by each doctor. Sets of easier/harder vignettes results in higher/lower doctor and algorithm accuracy scores. As the results for the expected disablement and expected sufficiency are almost identical, we show only the results for the expected sufficiency. Complete results are listed in appendix H.

The complete results for the individual doctors, posterior, expected disablement, and expected sufficiency ranking are included in Appendix H. Figure 5 compares the accuracy of each doctor to the associative and counterfactual algorithms. Each point gives the average accuracy for one of the 44 doctors, calculated as the proportion of diagnoses return by the doctor that contain the model disease. The first plot compares doctors to the associative (posterior ranking) algorithm. We refer to the set of vignettes considered by a single doctor as a

Agent	Accuracy (%)	$W_{\geq D}$	$W_{\geq A}$	$W_{\geq C1}$	$W_{\geq C2}$
D	71.98 ± 4.34	-	0.55 (0.25)	0.23 (0.11)	0.23 (0.11)
A	72.15 ± 4.30	0.55 (0.14)	-	0.07 (0)	0.07 (0)
C1	76.99 ± 4.03	0.81 (0.32)	1.0 (0.29)	-	-
C2	76.77 ± 3.83	0.80 (0.32)	1.0 (0.27)	-	-

TABLE II: **Group mean accuracy of doctors and algorithms**. The mean accuracy of the doctors D, associative A and counterfactual algorithms (C1 = expected sufficiency, C2 = expected disablement), averaged over all experiments. $W_{\geq K}$ gives the proportion of trials where this agent achieved a mean accuracy the same or higher than the mean accuracy of agent $K \in \{D, A, C1, C2\}$. The bracketed term is the proportion of trials where the agent scored the same or higher accuracy than agent K to 95% confidence, determined by a one sided binomial test. The two counterfactual measures are almost identical and so are not compared (see appendix H for comparison)

case set. There are roughly two types of performance for the doctors and algorithms, depending on the difficulty of the vignettes included in the case set. Doctors tend to achieve higher accuracies in case sets involving simpler vignettes—identified by high doctor and algorithm accuracies. Conversely, the algorithm tends to achieve higher accuracy than the doctors for case sets with more challenging vignettes—identified by low doctor and algorithm accuracies. This suggests that the diagnostic algorithms are complimentary to the doctors, with the algorithm performing particularly well on vignettes where doctor error is more common and vice versa. Overall, the associative algorithm performs on par with the average doctor, achieving a mean accuracy across all trails of $72.15 \pm 4.3\%$ v.s $71.98 \pm 4.34\%$ for doctors, meaning that on average we expect it to be in the top 49% of doctors. The algorithm scores higher than 20 of the doctors, draws with 4 of the doctors, and scores lower than 20 of the doctors.

The second graph in Figure 5 compares the expected sufficiency to the doctors over the same 44 trials. The separation between doctors and algorithm is now more pronounced, with the counterfactual algorithm detecting the model disease with higher accuracy than the majority of doctors. The counterfactual algorithm achieves a mean accuracy of $76.99 \pm 4.03\%$, considerably higher that the doctors and the associative algorithm and placing it in the top 20% of doctors in the cohort. The counterfactual algorithm scores higher than 34 of the doctors, draws with 2, and scores lower than 9. As with the associative algorithm, we observe a complementarity between the doctors and the algorithm, with the algorithm achieving high accuracies on case sets where the doctors achieved low accuracies and vice versa. In conclusion, the counterfactual algorithm outperforms both the associative algorithm and the average doctor in our cohort by a significant margin, achieving an expert clinical accuracy and achieving strong improvements for clinical vignettes that are difficult to diagnose or model rarer diseases.

IV. CONCLUSIONS

Poor access to primary healthcare and errors in primary diagnosis represent a significant challenge to global healthcare systems [1–4, 62, 63]. If Machine Learning assisted diagnosis is to help overcome these challenges, it is important that we first understand how human doctors perform diagnosis and to clearly define the desired output of our algorithms. Existing approaches have conflated diagnosis with associative inference. Whilst the former involves determining the underlying cause of a patient's symptoms, the latter involves learning correlations between patient data and disease occurrences, determining the most likely diseases in the population that the patient belongs to. This distinction places strong constraints on the accuracy of existing diagnostic algorithms. Overcoming these constraints requires that we fundamentally rethink how we define diagnosis and how we design diagnostic algorithms.

We have argued that diagnosis is fundamentally a causal inference task and presented a new causal definition of diagnosis. We have derived two counterfactual diagnostic measures that capture this causal definition, expected disablement and expected sufficiency, and a new class of diagnostic models—twin diagnostic networks—for calculating these measures. Using existing diagnostic models we have demonstrated that ranking disease hypotheses with these counterfactual measures greatly improves diagnostic accuracy compared to standard associative rankings. Whilst the associative algorithm performed on par with the average doctor in our cohort, the counterfactual algorithm performed better than 80% of the doctors—achieving expert clinical accuracy. Importantly, this improvement comes 'for free', without requiring any alterations to the disease model. Because of this backward compatibility our algorithm can be used as an immediate upgrade for existing Bayesian diagnostic algorithms including those outside of the medical setting [32–35, 64].

Whereas other approaches to improving diagnostic algorithms have focused on developing better model architectures [30] or exploiting new sources of data [65], our results demonstrate a new path towards expert-level diagnostic algorithms by employing causal and counterfactual reasoning to better mimic the decision making of human doctors. Indeed, our results add weight to the argument that machine learning methods that fail to incorporate causal reasoning will struggle to surpass the capabilities of human experts in certain domains [23]. Whilst we have focused on comparing our algorithms to doctors, future experiments could determine the effectiveness of these algorithms as clinical support systems—supporting doctors by providing a second opinion diagnosis. Given that our algorithm appears to be complimentary to human doctors, performing better on vignettes that doctors struggle to diagnose, it is likely that the combined diagnosis of doctor and algorithm could be significantly more accurate than either alone.

[1] H. Singh, G. D. Schiff, M. L. Graber, I. Onakpoya, and M. J. Thompson, "The global burden of diagnostic errors in primary care," *BMJ Qual Saf*, vol. 26, no. 6, pp. 484–494, 2017.

[2] M. L. Graber, "The incidence of diagnostic error in medicine," *BMJ Qual Saf*, vol. 22, no. Suppl 2, pp. ii21–ii27, 2013.

[3] H. Singh, T. D. Giardina, A. N. Meyer, S. N. Forjuoh, M. D. Reis, and E. J. Thomas, "Types and origins of diagnostic errors in primary care settings," *JAMA internal medicine*, vol. 173, no. 6, pp. 418–425, 2013.

[4] H. Singh, A. N. Meyer, and E. J. Thomas, "The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving us adult populations," *BMJ Qual Saf*, vol. 23, no. 9, pp. 727–731, 2014.

[5] D. R. Hogan, G. A. Stevens, A. R. Hosseinpoor, and T. Boerma, "Monitoring universal health coverage within the sustainable development goals: development and baseline data for an index of essential health services," *The Lancet Global Health*, vol. 6, no. 2, pp. e152–e168, 2018.

[6] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, *et al.*, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, p. 354, 2017.

[7] N. Brown and T. Sandholm, "Superhuman ai for multiplayer poker," *Science*, p. eaay2400, 2019.

[8] N. Tomašev, X. Glorot, J. W. Rae, M. Zielinski, H. Askham, A. Saraiva, A. Mottram, C. Meyer, S. Ravuri, I. Protsyuk, *et al.*, "A clinically applicable approach to continuous prediction of future acute kidney injury," *Nature*, vol. 572, no. 7767, p. 116, 2019.

[9] H. Liang, B. Y. Tsui, H. Ni, C. C. Valentim, S. L. Baxter, G. Liu, W. Cai, D. S. Kermany, X. Sun, J. Chen, *et al.*, "Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence," *Nature medicine*, p. 1, 2019.

[10] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature medicine*, vol. 25, no. 1, p. 44, 2019.

[11] J. De Fauw, J. R. Ledsam, B. Romera-Paredes, S. Nikolov, N. Tomasev, S. Blackwell, H. Askham, X. Glorot, *et al.*, "Clinically applicable deep learning for diagnosis and referral in retinal disease," *Nature medicine*, vol. 24, no. 9, p. 1342, 2018.

[12] K.-H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nature biomedical engineering*, vol. 2, no. 10, p. 719, 2018.

[13] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, "Artificial intelligence in healthcare: past, present and future," *Stroke and vascular neurology*, vol. 2, no. 4, pp. 230–243, 2017.

[14] G. Manogaran, V. Vijayakumar, R. Varatharajan, P. M. Kumar, R. Sundarasekar, and C.-H. Hsu, "Machine learning based big data processing framework for cancer diagnosis using hidden markov model and gm clustering," *Wireless personal communications*, vol. 102, no. 3, pp. 2099–2116, 2018.

[15] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature medicine*, vol. 25, no. 1, p. 24, 2019.

[16] R. Varatharajan, G. Manogaran, M. K. Priyan, and R. Sundarasekar, "Wearable sensor devices for early detection of alzheimer disease using dynamic time warping algorithm," *Cluster Computing*, vol. 21, no. 1, pp. 681–690, 2018.

[17] K. J. Kubota, J. A. Chen, and M. A. Little, "Machine learning for large-scale wearable sensor data in parkinson's disease: Concepts, promises, pitfalls, and futures," *Movement disorders*, vol. 31, no. 9, pp. 1314–1326, 2016.

[18] E. Rovini, C. Maremmani, and F. Cavallo, "How wearable sensors can support parkinson's disease diagnosis and treatment: a systematic review," *Frontiers in neuroscience*, vol. 11, p. 555, 2017.

[19] P. Croskerry, "From mindless to mindful practice, cognitive bias and clinical decision making," *New England Journal of Medicine*, vol. 368, no. 26, pp. 2445–2448, 2013.

[20] G. R. Norman, S. D. Monteiro, J. Sherbino, J. S. Ilgen, H. G. Schmidt, and S. Mamede, "The causes of errors in clinical reasoning: cognitive biases, knowledge deficits, and dual process thinking," *Academic Medicine*, vol. 92, no. 1, pp. 23–30, 2017.

[21] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[22] G. Marcus, "Deep learning: A critical appraisal," *arXiv preprint arXiv:1801.00631*, 2018.

[23] J. Pearl, "Theoretical impediments to machine learning with seven sparks from the causal revolution," *arXiv preprint arXiv:1801.04016*, 2018.

[24] H. L. Semigran, D. M. Levine, S. Nundy, and A. Mehrotra, "Comparison of physician and computer diagnostic accuracy," *JAMA internal medicine*, vol. 176, no. 12, pp. 1860–1861, 2016.

[25] R. A. Miller, M. A. McNeil, S. M. Challinor, F. E. Masarie Jr, and J. D. Myers, "The internist 1 quick medical reference project status report," *Western Journal of Medicine*, vol. 145, no. 6, p. 816, 1986.

[26] M. A. Shwe, B. Middleton, D. E. Heckerman, M. Henrion, E. J. Horvitz, H. P. Lehmann, and G. F. Cooper, "Probabilistic diagnosis using a reformulation of the internist-1/qmr knowledge base," *Methods of information in Medicine*, vol. 30, no. 04, pp. 241–255, 1991.

[27] R. Miller, "A history of the internist-1 and quick medical reference (qmr) computer-assisted diagnosis projects, with lessons learned," *Yearbook of medical informatics*, vol. 19, no. 01, pp. 121–136, 2010.

[28] D. E. Heckerman, E. J. Horvitz, and B. N. Nathwani, "Toward normative expert systems: Part i the pathfinder project," *Methods of information in medicine*, vol. 31, no. 02, pp. 90–105, 1992.

[29] D. E. Heckerman, E. J. Horvitz, and B. N. Nathwani, "Toward normative expert systems: Part i the pathfinder project," *Methods of information in medicine*, vol. 31, no. 02, pp. 90–105, 1992.

[30] S. Razzaki, A. Baker, Y. Perov, K. Middleton, J. Baxter, D. Mullarkey, D. Sangar, M. Taliercio, M. Butt, A. Majeed, *et al.*, "A comparative study of artificial intelligence and human doctors for the purpose of triage and diagno-

sis," *arXiv preprint arXiv:1806.10698*, 2018.
[31] M. Trimble and P. Hamilton, "The thinking doctor: clinical decision making in contemporary medicine," *Clinical Medicine*, vol. 16, no. 4, pp. 343–346, 2016.
[32] B. Cai, L. Huang, and M. Xie, "Bayesian networks in fault diagnosis," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 5, pp. 2227–2240, 2017.
[33] Z. Yongli, H. Limin, and L. Jinling, "Bayesian networks-based approach for power systems fault diagnosis," *IEEE Transactions on Power Delivery*, vol. 21, no. 2, pp. 634–639, 2006.
[34] S. Dey and J. Stori, "A bayesian network approach to root cause diagnosis of process variations," *International Journal of Machine Tools and Manufacture*, vol. 45, no. 1, pp. 75–91, 2005.
[35] B. Cai, Y. Liu, Q. Fan, Y. Zhang, Z. Liu, S. Yu, and R. Ji, "Multi-source information fusion based fault diagnosis of ground-source heat pump using bayesian network," *Applied energy*, vol. 114, pp. 1–9, 2014.
[36] J. Pearl, *Causality*. Cambridge university press, 2009.
[37] Q. Morris, "Recognition networks for approximate inference in bn20 networks," in *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pp. 370–377, Morgan Kaufmann Publishers Inc., 2001.
[38] E. Heckerman and N. Nathwani, "Toward normative expert systems: part ii probability-based representations for efficient knowledge acquisition and inference," *Methods of Information in medicine*, vol. 31, no. 02, pp. 106–116, 1992.
[39] J. Peters, D. Janzing, and B. Schölkopf, "Elements of causal inference-foundations and learning algorithms," 2017.
[40] C. M. Lee and R. W. Spekkens, "Causal inference via algebraic geometry: feasibility tests for functional causal structures with two binary observed variables," *Journal of Causal Inference*, vol. 5, no. 2.
[41] S. L. Lauritzen, *Graphical models*, vol. 17. Clarendon Press, 1996.
[42] G. H. Guyatt, C. Bombardier, and P. X. Tugwell, "Measuring disease-specific quality of life in clinical trials.," *CMAJ: Canadian Medical Association Journal*, vol. 134, no. 8, p. 889, 1986.
[43] M. P. Wellman and M. Henrion, "Explaining'explaining away'," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 3, pp. 287–292, 1993.
[44] J. Pearl, "Comment understanding simpson's paradox," *The American Statistician*, vol. 68, no. 1, pp. 8–13, 2014.
[45] S. Greenland, J. Pearl, J. M. Robins, *et al.*, "Causal diagrams for epidemiologic research," *Epidemiology*, vol. 10, pp. 37–48, 1999.
[46] I. Shpitser and J. Pearl, "Effects of treatment on the treated: Identification and generalization," in *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*, pp. 514–521, AUAI Press, 2009.
[47] S. L. Morgan and C. Winship, *Counterfactuals and causal inference*. Cambridge University Press, 2015.
[48] J. Pearl *et al.*, "Causal inference in statistics: An overview," *Statistics surveys*, vol. 3, pp. 96–146, 2009.
[49] J. Y. Halpern, *Actual causality*. MiT Press, 2016.
[50] D. Nikovski, "Constructing bayesian networks for medical diagnosis from incomplete and partially correct statistics," *IEEE Transactions on Knowledge & Data Engineering*, no. 4, pp. 509–516, 2000.
[51] I. Rish, M. Brodie, and S. Ma, "Accuracy vs. efficiency trade-offs in probabilistic diagnosis," in *AAAI/IAAI*, pp. 560–566, 2002.
[52] A. Oniśko, M. J. Druzdzel, and H. Wasyluk, "Learning bayesian network parameters from small data sets: Application of noisy-or gates," *International Journal of Approximate Reasoning*, vol. 27, no. 2, pp. 165–182, 2001.
[53] Y. Halpern and D. Sontag, "Unsupervised learning of noisy-or bayesian networks," *arXiv preprint arXiv:1309.6834*, 2013.
[54] A. Balke and J. Pearl, "Counterfactual probabilities: Computational methods, bounds and applications," in *Proceedings of the Tenth international conference on Uncertainty in artificial intelligence*, pp. 46–54, Morgan Kaufmann Publishers Inc., 1994.
[55] I. Shpitser and J. Pearl, "What counterfactuals can be tested," *arXiv preprint arXiv:1206.5294*, 2012.
[56] J. W. Peabody, J. Luck, P. Glassman, S. Jain, J. Hansen, M. Spell, and M. Lee, "Measuring the quality of physician practice by using clinical vignettes: a prospective validation study," *Annals of internal medicine*, vol. 141, no. 10, pp. 771–780, 2004.
[57] J. Veloski, S. Tai, A. S. Evans, and D. B. Nash, "Clinical vignette-based surveys: a tool for assessing physician practice variation," *American Journal of Medical Quality*, vol. 20, no. 3, pp. 151–157, 2005.
[58] L. Converse, K. Barrett, E. Rich, and J. Reschovsky, "Methods of observing variations in physicians' decisions: the opportunities of clinical vignettes," *Journal of general internal medicine*, vol. 30, no. 3, pp. 586–594, 2015.
[59] T. R. Dresselhaus, J. W. Peabody, J. Luck, and D. Bertenthal, "An evaluation of vignettes for predicting variation in the quality of preventive care," *Journal of General Internal Medicine*, vol. 19, no. 10, pp. 1013–1018, 2004.
[60] H. L. Semigran, J. A. Linder, C. Gidengil, and A. Mehrotra, "Evaluation of symptom checkers for self diagnosis and triage: audit study," *bmj*, vol. 351, p. h3480, 2015.
[61] K. Middleton, M. Butt, N. Hammerla, S. Hamblin, K. Mehta, and A. Parsa, "Sorting out symptoms: design and evaluation of the'babylon check'automated triage system," *arXiv preprint arXiv:1606.02041*, 2016.
[62] J. Higgs, M. A. Jones, S. Loftus, and N. Christensen, *Clinical Reasoning in the Health Professions E-Book*. Elsevier Health Sciences, 2008.
[63] A. L. Liberman and D. E. Newman-Toker, "Symptom-disease pair analysis of diagnostic error (spade): a conceptual framework and methodological approach for unearthing misdiagnosis-related harms using big data," *BMJ Qual Saf*, vol. 27, no. 7, pp. 557–566, 2018.
[64] C. Romessis and K. Mathioudakis, "Bayesian network approach for gas path fault diagnosis," *Journal of engineering for gas turbines and power*, vol. 128, no. 1, pp. 64–72, 2006.
[65] M. Rotmensch, Y. Halpern, A. Tlimat, S. Horng, and D. Sontag, "Learning a health knowledge graph from electronic medical records," *Scientific reports*, vol. 7, no. 1, p. 5994, 2017.
[66] D. Heckerman, "A tractable inference algorithm for diagnosing multiple diseases," in *Machine Intelligence and Pattern Recognition*, vol. 10, pp. 163–171, Elsevier, 1990.
[67] Y. Liu, K. Liu, and M. Li, "Passive diagnosis for wireless sensor networks," *IEEE/ACM Transactions on Networking (TON)*, vol. 18, no. 4, pp. 1132–1144, 2010.
[68] L. Perreault, S. Strasser, M. Thornton, and J. W. Shep-

pard, "A noisy-or model for continuous time bayesian networks.," in *FLAIRS Conference*, pp. 668–673, 2016.
[69] S. Arora, R. Ge, T. Ma, and A. Risteski, "Provable learning of noisy-or networks," in *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1057–1066, ACM, 2017.
[70] A. Abdollahi and K. Pattipati, "Unification of leaky noisy or and logistic regression models and maximum a posteriori inference for multiple fault diagnosis using the unified model," in *DX Conference (Denver, Co)*, 2016.
[71] J. Pearl, "Probabilities of causation: three counterfactual interpretations and their identification," *Synthese*, vol. 121, no. 1-2, pp. 93–149, 1999.
[72] I. Shpitser and J. Pearl, "What counterfactuals can be tested," in *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pp. 352–359, AUAI Press, 2007.
[73] J. Y. Halpern and J. Pearl, "Causes and explanations: A structural-model approach. part i: Causes," *The British journal for the philosophy of science*, vol. 56, no. 4, pp. 843–887, 2005.
[74] J. Y. Halpern, "Axiomatizing causal reasoning," *Journal of Artificial Intelligence Research*, vol. 12, pp. 317–337, 2000.
[75] T. Eiter and T. Lukasiewicz, "Complexity results for structure-based causality," *Artificial Intelligence*, vol. 142, no. 1, pp. 53–89, 2002.
[76] In this context, a model or algorithm is interpretable if it is possible to determine why the algorithm has reached a given diagnosis
[77] equivalent to board certified primary care physicians
[78] 2: It should be noted that the disease model evaluated in the following experiments is not the current production model used for the purposes of diagnosis and triage by Babylon Health$^{\text{TM}}$. This article is for general information and academic purposes, and this disease model is used to facilitate discussion on this topic. This article is not designed to be relied upon for any other purpose.

Appendices

The structure of these appendices is as follows. In appendix A we detail our notation. In appendix B we outline the tools we use to derive our results – namely the frameworks of structural causal models (SCMs), introduce noisy-or Bayesian networks, and derive their SCM representation. In appendix C we outline the framework of twin-networks [54], and derive a simplified class of twin networks that we will use for computing our counterfactual diagnostic measures ('twin diagnostic networks'). In appendices D and F we introduce and derive expressions for our counterfactual diagnostic measure s—the expected sufficiency and the expected disablement—for the family of noisy-or diagnostic networks introduced in Appendices B and C. In appendices E and G we prove that these two measures satisfy our desiderata. In appendix H we list our experimental results.

A. NOTATION

Variables: For the disease models we consider, all variables X are Bernoulli, $X \in \{0, 1\}$. Where appropriate we refer to $X = 0$ as the variable X being 'off', and $X = 1$ as the variable X being 'on'. We denote single variables as capital Roman letters, and sets of variables as calligraphic, e.g. $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$. The union of two sets of variables $\mathcal{X}$ and $\mathcal{Y}$ is denoted $\mathcal{X} \cup \mathcal{Y}$, the intersection is denoted $\mathcal{X} \cap \mathcal{Y}$, and the relative compliment of $\mathcal{X}$ w.r.t $\mathcal{Y}$ as $\mathcal{X} \setminus \mathcal{Y}$. The instantiation of a single variable is indicated by a lower case letter, $X = x$, and for a set of variables $\mathcal{X} = \underline{x}$ denotes some arbitrary instantiation of all variables belonging to $\mathcal{X}$, e.g. $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$. The probability of $\mathcal{X} = \underline{x}$ is denoted $p(\mathcal{X} = \underline{x})$, and sometimes for simplicity is denoted as $p(\underline{x})$.

For a given variable X and a directed acyclic graph (DAG) G, we denote the set of parents of X as $\mathsf{Pa}(X)$, the set of children of X as $\mathsf{Ch}(X)$, all ancsertors of X as $\mathsf{Anc}(X)$, and all descendent of X as $\mathsf{Dec}(X)$. if we perform a graph cut operation on G, removing a directed edge from Y to X, we denote the variable X in the new DAG generated by this cut as $X^{\setminus Y}$.

Functions: Bernoulli variables are represented interchangeably as Boolean variables, with $1 \leftrightarrow$ 'True' and $0 \leftrightarrow$ 'False'. For a given instantiation of a Bernoulli/Boolean variable $X = x$, we denote the negation of x as $\bar{x}$ – for example if $x = 1(0)$, $\bar{x} = 0(1)$. We denote the Boolean AND function as $\wedge$, and the Boolean OR function as $\vee$.

B. STRUCTURAL CAUSAL MODELS

First we define structure causal models (SCMs), sometimes also called structural equation models or functional causal models. These are widely applied and studied probabilistic models, and their relation to other approaches such

as Bayesian networks are well understood [36, 41]. The key characteristic of SCMs is that they represent variables as functions of their direct causes, along with an exogenous 'noise' variable that is responsible for their randomness.

Definition 3 (Structural Causal Model)**.** *A causal model specifies:*

1. *a set of latent, or* noise*, variables $U = \{u_1, \ldots, u_n\}$, distributed according to $P(U)$.*
2. *a set of observed variables $V = \{v_1, \ldots, v_n\}$,*
3. *a directed acyclic graph G, called the* causal structure *of the model, whose nodes are the variables $U \cup V$,*
4. *a collection of functions $F = \{f_1, \ldots, f_n\}$, where f_i is a mapping from $U \cup V/v_i$ to v_i. The colelction F forms a mapping from U to V. This is symbolically represented as*

$$v_i = f_i(pa_i, u_i), \text{ for } i = 1, \ldots, n,$$

where pa_i denotes the parent nodes of the ith observed variable in G.

As the collection of functions F forms a mapping from noise variables U to observed variables V, the distribution over noise variables induces a distribution over observed variables, given by

$$P(v_i) := \sum_{u | v_i = f_i(\mathrm{pa}_i, u)} P(u), \text{ for } i = 1, \ldots, n. \tag{6}$$

We can hence assign uncertainty over observed variables despite the the underlying dynamics being deterministic.

In order to formally define a counterfactual query, we must first define the interventional primitive known as the "*do*-operator" [36]. Consider a SCM with functions F. The effect of intervention $do(X = x)$ in this model corresponds to creating a new SCM with functions $F_{X=x}$, formed by deleting from F all functions f_i corresponding to members of the set X and replacing them with the set of constant functions $X = x$. That is, the *do*-operator forces variables to take certain values, regardless of the original causal mechanism. This represents the operation whereby an agent intervenes on a variable, fixing it to take a certain value. Probabilities involving the *do*-operator, such as $P(Y = y|do(X = x))$, correspond to evaluating ordinary probabilities in the SCM with functions F_X, in this case $P(Y = y)$. Where appropriate, we use the more compact notation of Y_x to denote the variable Y following the intervention $\mathrm{do}(X = x)$.

Next we define noisy-OR models, a specific class of SCMs for Bernoulli variables that are widely employed as diagnostic models [26, 27, 33, 50, 53, 66–70]. The noisy-OR assumption states that a variable Y is the Boolean OR of its parents $X_1, X_2, \ldots, X_n$, where the inclusion or exclusion of each causal parent in the OR function is decided by an independent probability or 'noise' term. The standard approach to defining noisy-OR is to present the conditional independence constraints generated by the noisy-OR assumption [71],

$$p(Y = 0 \,|\, X_1, \ldots, X_n) = \prod_{i=1}^{n} p(Y = 0 \,|\, \mathrm{only}(X_i = 1)) \tag{7}$$

where $p(Y = 0 \,|\, \mathrm{only}(X_i = 1))$ is the probability that $Y = 0$ conditioned on all of its (endogenous) parents being 'off' ($X_j = 0$) except for X_i alone. We denote $p(Y = 0 \,|\, \mathrm{only}(X_i = 1)) = \lambda_{X_i,Y}$ by convention.

The utility of this assumption is that it reduces the number of parameters needed to specify a noisy-OR network to $\mathcal{O}(N)$ where N is the number of directed edges in the network. All that is needed to specify a noisy-OR network are the single variable marginals $p(X_i = 1)$ and, for each directed edge $X_i \to Y_j$, a single λ_{X_i,Y_j}. For this reason, noisy-OR has been a standard assumption in Bayesian diagnostic networks, which are typically large and densely connected and so could not be efficiently learned and stored without additional assumptions on the conditional probabilities. We now define the noisy-OR assumption for SCMs.

Definition 4 (noisy-OR SCM)**.** *A noisy-OR network is an SCM of Bernoulli variables, where for any variable Y with parents $Pa(Y) = \{X_1, \ldots, X_N\}$ the following conditions hold*

1. *Y is the Boolean OR of its parents, where for each parent X_i there is a Bernoulli variable U_i whose state determines if we include that parent in the OR function or not*

$$y = \bigvee_{i=1}^{N} (x_i \wedge \bar{u}_i) \tag{8}$$

i.e. $Y = 1$ *if any parent is on,* $x_i = 1$*, and is not ignored,* $u_i = 0$ *(*$\bar{u}_i = 1$ *where 'bar' denotes the negation of* u_i*).*

2. *The exogenous latent encodes the likelihood of ignoring the state of each parent in (1),* $P(u_Y) = p(u_1, u_2, \ldots, u_N)$*. The probability of ignoring the state of a given parent variable is independent of whether you have or have not ignored any of the other parents,*

$$P(u_1, u_2, \ldots, u_N) = \prod_{i=1}^{N} P(u_i)$$

3. *For every node* Y *their is a parent 'leak node'* L_Y *that is singly connected to* Y *and is always 'on', with a probability of ignoring given by* λ_{L_Y}

The leak node (assumption 3) represents the probability that $Y = 1$, even if $X_i = 0 \; \forall \; X_i \in \text{Pa}(Y)$. This allows $X_i = 1$ to be caused by an exogenous factor (outside of our model). For example, the leak nodes allow us to model the situation that a disease spontaneously occurs, even if all risk factors that we model are absent, or that a symptom occurs but none of the diseases that we model have caused it. It is conventional to treat the leak node associated with a variable Y as a parent node L_Y with $p(L_Y = 1)$. Every variable in the noisy-OR SCM has a single, independent leak node parent.

Given Definition 4, why is the noisy-or assuption justified for modelling diseases? First, consider the assumption (1), that the generative function is a Boolean OR of the individual parent 'activation functions' $x_i \cap \bar{u}_i$. This is equivalent to assuming that the activations from diseases or risk-factors to their children never 'destructively interfere'. That is, if D_i is activating symptom S, and so is D_j, then this joint activation never cancels out to yield $S = F$. As a consequence, all that is required for a symptom to be present is that at least one disease to be causing it, and likewise for diseases being caused by risk factors. This property of noisy-OR, whereby an individual cause is also a sufficient cause, is a natural assumption for diseases modelling – where diseases are (typically by definition) sufficient causes of their symptoms, and risk factors are defined such that they are sufficient causes of diseases. For example, if preconditions $R_1 = 1$ and $R_2 = 1$ are needed to cause $D = 1$, then we can represent this as a single risk factor $R = R_1 \wedge R_2$. Assumption 2 states that a given disease (risk factor) has a fixed likelihood of activating a symptom (disease), independent of the presence or absence of any other disease (risk factor). In the noisy-or model, the likelihood that we ignore the state of a parent X_i of variable Y_i is given by

$$p(u_i = 1) = \frac{p(Y_i = 0 |\ \text{do}(X_i = 1))}{p(Y_i = 0 |\ \text{do}(X_i = 0))} \tag{9}$$

and so is directly associated with a (causal) relative risk. In the case that child Y has two parents, X_1 and X_2, noisy-OR assumes that this joint relative risk factorises as

$$p(u_1 = 1, u_2 = 1) = \frac{p(Y = 0 |\ \text{do}(X_1 = 1, X_2 = 1))}{p(Y = 0 |\ \text{do}(X_1 = 0, X_2 = 0))} = \frac{p(Y = 0 |\ \text{do}(X_1 = 1))}{p(Y = 0 |\ \text{do}(X_1 = 0))} \times \frac{p(Y = 0 |\ \text{do}(X_2 = 1))}{p(Y = 0 |\ \text{do}(X_2 = 0))} = p(u_1 = 1)p(u_2 = 1) \tag{10}$$

Whilst it is likely that interactions between causal parents will mean that these relative risks are not always multiplicative, it is assumed to be a good approximation. For example, we assume that the likelihood that a disease fails to activate a symptoms is independent of whether or not any other disease similarly fails to activate that symptom.

As noisy-OR models are typically presented as Bayesian networks, the above definition of noisy-OR is non-standard. We now show that the SCM definition yields the Bayesian network definition, (7).

Theorem 3 (noisy-OR CPT)**.** *The conditional probability distribution of a child* Y *given its parents* $\{X_1, \ldots, X_n\}$ *and obeying Definition 14 is given by*

$$p(Y = 0 \,|\, X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} \lambda_{X_i,Y}^{x_i} \tag{11}$$

where

$$\lambda_{X_i,Y} = p(Y = 0 |\ \textit{only}\ (X_i = 1)) \tag{12}$$

Proof. For $Y = 0$, the negation of y, denoted $\bar{y}$, is given by

$$\bar{y} = \neg \left(\bigvee_{i=1}^{N} (x_i \wedge \bar{u}_i) \right) = \bigwedge_{i=1}^{N} (\bar{x}_i \vee u_i) \tag{13}$$

The CPT is calculated from the structural equations by marginalizing over the latents. I.e. we sum over all latent states that yield $Y = 0$. Equivalently, we can marginalize over all exogenous latent states multiplied by the above Boolean function, which is 1 if the condition $Y = 0$ is met, and 0 otherwise.

$$\begin{aligned}
p(Y = 0 \,|\, X_1 = x_1, \ldots, X_n = x_n) &= \sum_{u_Y} \bigwedge_{i=1}^{N} (\bar{x}_i \vee u_i)\, p(u_Y) \\
&= \sum_{u_{X_i,Y}} \prod_{X_i} (\bar{x}_i \vee u_i) \prod_{U_{X_i,Y}} p(u_{X_i,Y}) \\
&= \prod_{X_i} \sum_{u_{X_i,Y}} p(u_{X_i,Y})\, (\bar{x}_i \vee u_i) \\
&= \prod_{X_i} [p(u_{X_i,Y} = 1) + p(u_{X_i,Y} = 0)\bar{x}_i] \\
&= \prod_{X_i} [\lambda_{X_i,Y} + (1 - \lambda_{X_i,Y})\bar{x}_i] \\
&= \prod_{X_i} \lambda_{X_i,Y}^{x_i}
\end{aligned} \tag{14}$$

This is identical to the noisy-OR cpt (7) □

The leak node is included as a parent x_L where $p(x_L = 1) = 1$, and a (typically large) probability of being ignored λ_L. This node represents the likelihood that Y will be activated by some causal influence outside of the model, and is included to ensure that $p(Y = 1 | \wedge_{i=1}^{n} (X_i = 0)) \neq 0$. As the leak node is always on, its notation can be suppressed and it is standard notation to write the CPT as

$$p(Y = 0 \,|\, X_1 = x_1, \ldots, X_n = x_n) = \lambda_L \prod_{X_i} \lambda_{X_i,Y}^{x_i} \tag{15}$$

C. TWIN DIAGNOSTIC NETWORKS

In this appendix we derive the structure of diagnostic twin networks. First we provide a brief overview to the twin-networks approach to counterfactual inference. See [54] and [72] for more details on this formalism. First, recalling the definition of the *do* operator from the previous section, we define counterfactuals as follows.

Definition 5 (Counterfactual)**.** *Let X and Y be two subsets of variables in V. The counterfactual sentence "Y would be y (in situation U), had X been x," is the solution $Y = y$ of the set of equations F_x, succinctly denoted $Y_x(U) = y$.*

As with observed variables in Definition 3, the latent distribution $P(U)$ allows one to define the probabilities of counterfactual statements in the same manner they are defined for standard probabilities (6).

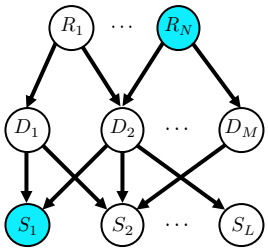
$$P(Y_x = y) = \sum_{u | Y_x(u) = y} P(u). \tag{16}$$

Reference [36] provides an algorithmic procedure for computing arbitrary counterfactual probabilities for a given SCM. First, the the distribution over latents is updated to account for the observed evidence. Second, the *do*-operator is applied, representing the counterfactual intervention. Third, the new causal model created by the application of the *do*-operator in the previous step is combined with the updated latent distribution to compute the counterfactual query. In general, denote $\mathcal{E}$ as the set of factual evidence. The above can be summarised as,

1. (abduction). The distribution of the exogenous latent variables $P(u)$ is updated to obtain $P(u \mid \mathcal{E})$
2. (action). Apply the do-operation to the variables in set X, replacing the equations $X_i = f_i(\mathrm{Pa}(x_i), u_i)$ with $X_i = x_i \; \forall \; X_i \in X$.
3. (prediction). Use the modified model to compute the probability of $Y = y$.

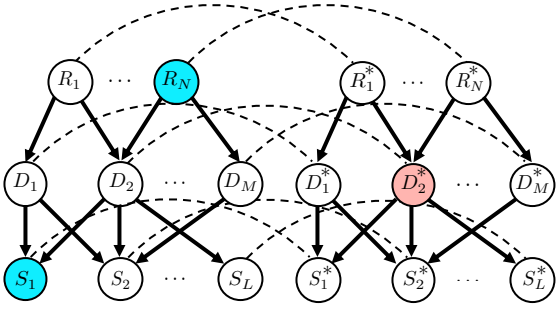
The issue with applying this approach to our large diagnostic models is that the first step, updating the exogenous latents, is in general intractable for models with large tree-width. The twin-networks formalism, introduced in [54], is a method for greatly reduces and amortises the cost of this procedure. Rather than explicitly updating the exogenous latents, performing an intervention, and performing belief propagation on the resulting SCM, twin networks allow us to calculate the counterfactual by performing belief propagation on a single 'twin' SCM – without requiring the expensive abduction step. The twin network is constructed as a composite of two copies of the original SCM where copied variables share their corresponding latents [54]. We refer to pairs of copied variables as 'dual variables'. Nodes on this twin network can then be merged following simple rules outlined in [72], further reducing the complexity of computing the counterfactual query. We now outline the process of constructing the twin diagnostic network in the case of the two counterfactual queries we are interested in – those with single counterfactual interventions, and those where all counterfactual variables bar one are intervened on.

[A]



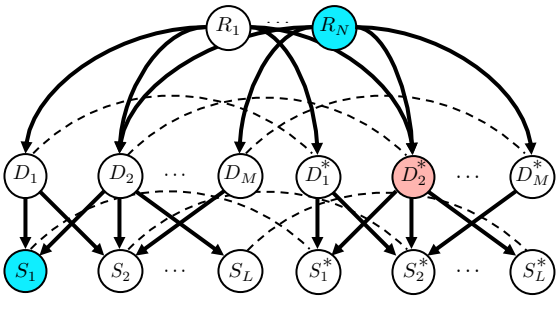
Factual graph

[B]



Factual graph Counterfactual graph

[C]



Factual graph Counterfactual graph

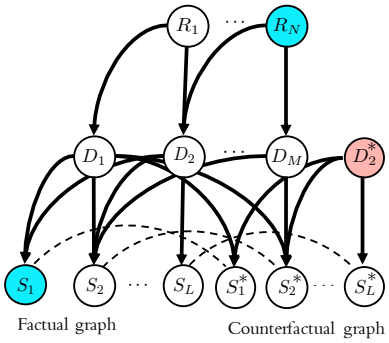
We assume the DAG structure of our diagnostic model is a three layer network [A]. The top layer nodes represent risk factors, the second layer represent diseases, and the third layer symptoms. We assume no directed edges between nodes belonging to the same layer. To construct the twin network, first the SCM in [A] is copied. In [B] the network

on the left will encode the factual evidence in our counterfactual query, and we refer to this as the factual graph. The network on the right in [B] will encode our counterfactual interventions and observations, and we refer to this as the counterfactual graph. We use an asterisk X^* to denote the counterfactual dual variable of X.

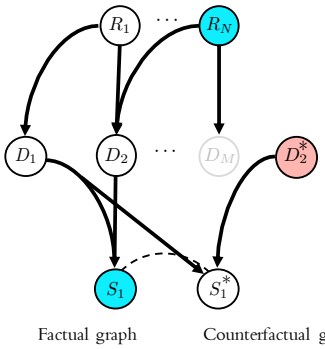
As detailed in [54], the twin network is constructed such that each node on the factual graph shares its exogenous latent with its dual node, so $u^*_{X_i} = u_{X_i}$. These shared exogenous latents are shown as dashed lines in figures [B-E]. First, we consider the case where we perform a counterfactual intervention on a single disease. As shown in [B], we select a disease node in the counterfactual graph to perform our intervention on (in this instance D^*_2). In Figure [C], blue circles represent observations and red circles represent interventions. The do-operation severs any directed edges going into D^* and fixes $D^* = 0$, as shown in [D] below.

Once the counterfactual intervention has been applied, it is possible to greatly simplify the twin network graph structure via node merging [72]. In SCM's a variable takes a fixed deterministic value given an instantation of all of its parents and its exogenous latent. Hence, if two nodes have identical exogenous latents and parents, they are copies and can be merged into a single node. By convention, when we merge these identical dual nodes we map $X^* \mapsto X$ (dropping the asterisk). Dual nodes which share no ancestors that have been intervened upon can therefore be merged. As we do not perform interventions on the risk factor nodes, all (R_i, R^*_i) are merged (note that for the sake of clarity we do not depict the exogenous latents for risk factors).

[D]



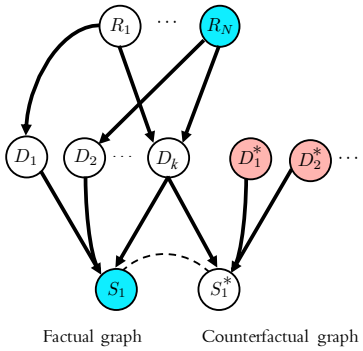
[E]



Next, we merge all dual factual/counterfactual disease nodes that are not intervened on, as their latents and parents are identical (shown in [D]). Finally, any symptoms that are not children of the disease we have intervened upon can be merged, as all of their parent variables are identical. The resulting twin network is shown in [E]. Note that we have also removed any superfluous symptom nodes that are unevidenced, as they are irrelevant for the query.

In the case that we intervene on all of the counterfactual diseases except one, following the node merging rule outlined above, we arrive at a model with a single disease that is a parent of both factual and counterfactual symptoms, as shown in Figure [F].

[F]



We refer to the SCMs shown in figures [E] and [F] as 'twin diagnostic networks'. The counterfactual queries we are interested in can be determined by applying standard inference techniques like belief propagation to these models.

D. EXPECTED SUFFICIENCY

In this appendix we derive a simple closed form expression for our proposed diagnostic measure, the *expected sufficiency*, which corresponds to the case where we perform counterfactual interventions on all diseases bar one (D_k, model shown in Figure [F]). We derive our expressions for three layer noisy-OR SCM's. Before proceeding, we motivate our choice of counterfactual query for the task of diagnosis.

An observation will often have multiple possible causes, which constitute competing explanations. For example, the observation of a symptom $S = 1$ can in principle be explained by any of its parent diseases. In the case that a symptom has multiple associated causes (diseases), rarely is a single disease *necessary* to explain a given symptom. Equivalently, the symptoms associated with a disease tend to be present in patient's suffering from this diseases, without *requiring* a secondary disease to be present. This can be summarised by the following assumption – *any single disease is a sufficient cause of any of its associated symptoms*. Under this assumption, determining the likelihood that a diseases is causing a symptom reduces to simple deduction – removing all other possible causes and seeing if the symptom remains.

The question of how we can define and quantify causal explanations in general models is an area of active research [49, 73–75] and the approach we propose here cannot be applied to all conceivable SCMs. For example, if you had a symptom that can be present *only if* two parents diseases D_1 and D_2 are both present, then neither of these parents in isolation is a sufficient cause (individually, $D_1 = 1$ and $D_2 = 1$ are necessary but not sufficient to cause $S = 1$). In Appendix A.4 we present a different counterfactual query that captures causality in this case by reasoning about necessary treatments. However, in the case that our symptoms obey noisy-or statistics, all diseases are *individually sufficient* to generate any symptom. This is ensured by the OR function, which states that a symptom S is the Boolean OR of its parents individual activation functions, $s = \bigvee_{i=1}^{N}[d_i \wedge \bar{u}_{D_i,S}]$ where the activation function from parent D_i is $f_i = d_i \wedge \bar{u}_{D_i,S}$. Thus, any single activation is sufficient to explain $S = 1$ and we can quantify the expected sufficiency of a diseases individually. An example of a model that would violate this property is a noisy-AND model, where $s = \bigwedge_{i=1}^{N}[d_i \wedge \bar{u}_{D_i,S}]$ - e.g. all parent diseases must be present in order for the symptom to be present.

Given these properties of noisy-OR models (as disease models in general), we propose our measure for quantifying how well a disease explains the patient's symptoms – the *expected sufficiency*. For a given disease, this measures the number of symptoms that we would expect to remain if we intervened to nullify all other possible causes of symptoms. This counterfactual intervention is represented by the causal model shown in figure [F] in appendix A.2.

Definition 2. *The expected sufficiency of disease D_k determines the number of positively evidenced symptoms we would expect to persist if we intervene to switch off all other possible causes of the symptoms,*

$$\mathbb{E}_{suff}(D_K, \mathcal{E}) := \sum_{\mathcal{S}'} \left|\mathcal{S}'_+\right| p(\mathcal{S}'|\mathcal{E}, do(\mathcal{D} \setminus D_k = 0), do(\mathcal{U}_L = 0)) \tag{17}$$

where $\mathcal{D} \setminus D_k$ denotes the set of all diseases except for D_k, and $\mathcal{U}_L$ denotes the set of all symptom leak nodes. The sum is calculated over all possible counterfactual symptom states $\mathcal{S}'$, where $\mathcal{S}'_+$ denotes the set of positive symptoms in the counterfactual symptom state. $\mathcal{E}$ denotes the set of all factual evidence.

To evaluate the expected sufficiency we must first determine the form our the dual symptom CPTs in the corresponding twin network (figure [F]).

Lemma 1. *For a given symptom S and its counterfactual dual S^*, with parent diseases $\mathcal{D}$ and under the counterfactual interventions $do(\mathcal{D} \setminus D_k^* = 0)$ and $do(\mathcal{U}_L^* = 0)$, the joint conditional distribution is given by*

$$p(s, s^* | \wedge_{i=1}^N d_i, do(\wedge_{i\neq k} D_i^* = 0), do(u_L^* = 0)) = \begin{cases} p(s=0 | \wedge_{i=1}^N d_i), & s = s^* = 0 \\ 0, & s = 0, s^* = 1 \\ \lambda_{D_k,s}^{d_k} p(s^{\backslash k} = 1 | \wedge_{i\neq k} d_i, d_k = 1), & s = 1, s^* = 0 \\ (1 - \lambda_{D_k,S})\delta(d_k - 1), & s = 1, s^* = 1 \end{cases}$$

where $\delta(d_k - 1) = 1$ if $D_k = 1$ else 0, and $\underline{d}$ is an instantiation of all $D_i \in \mathsf{Pa}(S)$, $\wedge_{i\neq k} D_i^*$ is the set of all counterfactual disease nodes excluding D_k, $\wedge_{i\neq k} d_i$ is the given instantiation on all disease nodes exlcuding D_k, and u_L^* denotes the leak node for the counterfactual symptom. $s^{\backslash k}$ denotes the state of the factual symptom node S under the graph surgery removing any direct edge from D_k to S.

Proof. The CPT for the dual symptom nodes S, S^* is given by

$$\begin{aligned} &p(s, s^* | \underline{d}, \mathrm{do}(\wedge_{i\neq k} D_i^* = 0), \mathrm{do}(u_L^* = 0)) = \\ &\sum_{u_{D_1,S}} p(u_{D_1,S}) \cdots \sum_{u_{D_N,S}} p(u_{D_N,S}) \sum_{u_L} p(u_L) p(s | d_k, \wedge_{i\neq k} d_i, u_L) p(s^* | d_k, \mathrm{do}(\wedge_{i\neq k} D_i^* = 0), \mathrm{do}(u_L^* = 0)) \end{aligned} \tag{18}$$

Where we have use the fact that the latent variables and the disease variables together form a Markov blanket for S, S^*, and we have used the conditional independence structure of the twin network, shown in Figure [F], which implies that S and S^* only share a single variable, D_k, in their Markov blankets. With the full Markov blanket specified, including the exogenous latents, the CPTs in (18) are deterministic functions, each taking the value 1 if their conditional constraints are satisfied. Note that the product of these two functions is equivalent to a function that is 1 if both sets of conditional constraints are satisfied and zero otherwise, and marginalizing over all latent variable states multiplied by this function is equivalent to the definition of the CPT for SCMs given in equation (6), where the CPT is determined by a conditional sum over the exogenous latent variables. Given the definition of the noisy-OR SCM in (8), these functions take the form

$$p(s | d_k, \wedge_{i\neq k} d_i, u_L) = \begin{cases} \bar{u}_L \bigwedge_{i=1}^N [\bar{d}_i \vee u_{D_i,S}], & s = 0 \\ 1 - \bar{u}_L \bigwedge_{i=1}^N [\bar{d}_i \vee u_{D_i,S}], & s = 1 \end{cases} \tag{19}$$

and

$$p(s^* | d_k, \mathrm{do}(\wedge_{i\neq k} D_i^* = 0), \mathrm{do}(u_L^* = 0)) = \begin{cases} \bar{d}_k \vee u_{D_k,S}, & s = 0 \\ 1 - \bar{d}_k \vee u_{D_k,S}, & s = 1 \end{cases} \tag{20}$$

Taking the product of these functions gives the function $g_{s,s^*}(\underline{u}, \underline{d}, u_L) := p(s | d_k, \wedge_{i\neq k} d_i, u_L) \times p(s^* | d_k, \mathrm{do}(\wedge_{i\neq k} D_i^* = 0), \mathrm{do}(u_L^* = 0))$ where $\underline{u}$ denotes a given instantiation of the free latent variables $u_{D_1,S}, \ldots, u_{D_N,S}$.

$$g_{s,s^*}(\underline{u}, \underline{d}, u_L) = \begin{cases} \bar{u}_L \bigwedge\limits_{i=1}^N [\bar{d}_i \vee u_{D_i,S}], & s = s^* = 0 \\ 0, & s = 0, s^* = 1 \\ [\bar{d}_k \vee u_{D_k,S}] \wedge [1 - \bigwedge\limits_{i=1}^N [\bar{d}_i \vee u_{D_i,S}]], & s = 1, s^* = 0 \\ 1 - \bar{d}_k \vee u_{D_k,S}, & s = 1, s^* = 1 \end{cases} \tag{21}$$

$$p(s, s^*|\underline{d}, \mathrm{do}(\wedge_{i\neq k} D_i^* = 0), \mathrm{do}(u_L^* = 0)) = \sum_{u_{D_1,S}} p(u_{D_1,S}) \cdots \sum_{u_{D_N,S}} p(u_{D_N,S}) \sum_{u_L} p(u_L) g_{s,s^*}(\underline{u}, \underline{d}, u_L) \tag{22}$$

$$= \begin{cases} \lambda_L \prod_{i=1}^{N} \lambda_{D_i,S}^{d_i}, & s = s^* = 0 \\ 0, & s = 0, s^* = 1 \\ \lambda_{D_k,s}^{d_k} - \lambda_L \prod_{i=1}^{N} \lambda_{D_i,S}^{d_i}, & s = 1, s^* = 0 \\ (1 - \lambda_{D_k,S})\delta(d_k - 1), & s = 1, s^* = 1 \end{cases} \tag{23}$$

where we have used $\sum_{u_{D_i,S}} p(u_{D_i,S})\bar{d}_i \vee u_{D_i,S} = p(u_{D_i,S} = 1) + p(u_{D_i,S} = 0)\bar{d}_i = p(u_{D_i,S} = 1)^{d_i} = \lambda_{D_i,S}^{d_i}$, and $\sum_{u_{D_k,S}} p(u_{D_k,S})[1 - \bar{d}_k \vee u_{D_k,S}] = (1 - \lambda_{D_k,S})\delta(d_k - 1)$, where $\delta(d_k - 1)$ is 1 iff $d_k = 1$ and 0 otherwise. $\lambda_L \prod_{i=1}^{N} \lambda_{D_i,S}^{d_i}$ can immediately be identified as $p(s = 0|\mathcal{D})$ by (15). $\lambda_{D_k,s}^{d_k} - \lambda_L \prod_{i=1}^{N} \lambda_{D_i,S}^{d_i} = \lambda_{D_k,s}^{d_k}(1 - \lambda_L \prod_{i\neq k} \lambda_{D_i,S}^{d_i})$, and we can identify $\lambda_L \prod_{i\neq k} \lambda_{D_i,S}^{d_i} = p(s = 0| \wedge_{i\neq k} d_i, d_k = 0)$. Therefore $\lambda_{D_k,s}^{d_k} - \lambda_L \prod_{i=1}^{N} \lambda_{D_i,S}^{d_i} = \lambda_{D_k,s}^{d_k} p(s = 1| \wedge_{i\neq k} d_i, d_k = 0)$. Finally, we can express this as $\lambda_{D_k,s}^{d_k} p(s^{\backslash k} = 1| \wedge_{i\neq k} d_i, d_k = 1)$, where $s^{\backslash k}$ is the instantiation of $S^{\backslash k}$ – which is the variable generated by removing any directed edge $D_k \to S$ (or equivalently, replacing $\lambda_{D_k,S}$ with 1).

□

Given our expression for the symptom CPT on the twin network, we now derive the expression for the expected sufficiency.

Theorem 2. *For noisy-OR networks described in Appendix A.1-A.4, the expected sufficiency of disease D_k is given by*

$$\mathbb{E}_{suff}(D_K, \mathcal{E}) = \frac{1}{p(\mathcal{S}_\pm|\mathcal{R})} \sum_{\mathcal{S}\subseteq\mathcal{S}_+} |\mathcal{S}_+ \setminus \mathcal{S}| p(\mathcal{S}_- = 0, \mathcal{S}^{\backslash k} = 1, D_k = 1|\mathcal{R}) \prod_{s\in\mathcal{S}_+\setminus\mathcal{S}} (1 - \lambda_{k,s}) \prod_{s\in\mathcal{S}} \lambda_{k,s}$$

where $\mathcal{S}_\pm$ denotes the positive and negative symptom evidence, $\mathcal{R}$ denotes the risk-factor evidence, and $\mathcal{S}^{\backslash k}$ denotes the set of symptoms $\mathcal{S}$ with all directed arrows from D_k to $S \in \mathcal{S}$ removed.

Proof. Starting from the definition of the expected sufficiency

$$\mathbb{E}_{\text{suff}}(D_K, \mathcal{E}) := \sum_{\mathcal{S}'} \left|\mathcal{S}'_+\right| p(\mathcal{S}'|\mathcal{E}, \mathrm{do}(\mathcal{D} \setminus D_k = 0), \mathrm{do}(\mathcal{U}_L = 0)) \tag{24}$$

we must find expressions for all CPTs $P(\mathcal{S}'|\mathcal{E}, \mathrm{do}(\mathcal{D} \setminus D_k = 0), \mathrm{do}(\mathcal{U}_L = 0)$ where $|\mathcal{S}'_+| \neq 0$ (terms with $\mathcal{S}'_+ = \emptyset$ do not contribute to (24)). Lets $\{S \text{ s.t. } S \in \mathcal{S}_-, S^* \in \mathcal{S}'_-\}$ (symptoms that remain off following the counterfactual intervention), $\mathcal{S}_B = \{S \text{ s.t. } S \in \mathcal{S}_+, S^* \in \mathcal{S}'_+\}$ (symptoms that remain on following the counterfactual intervention), and $\mathcal{S}_C = \{S \text{ s.t. } S \in \mathcal{S}_+, S^* \in \mathcal{S}'_-\}$ (symptoms that are switched off by the counterfactual intervention). Lemma 56 implies that $p(S = 0, S^* = 1|\underline{d}, \mathrm{do}(\wedge_{i\neq k} D_i^* = 0), \mathrm{do}(u_L^* = 0)) = 0$, and therefore these three cases are sufficient to characterise all possible counterfactual symptom states $\mathcal{S}'$. Therefore, to evaluate (24), we need only determine expressions for the following terms

$$p(S_A^* = 0, S_B^* = 1, S_C^* = 0|\mathcal{S}_\pm, \mathcal{R}, \mathrm{do}(\wedge_{i\neq k} D_i^* = 0), \mathrm{do}(\mathcal{U}_L^* = 0)) \tag{25}$$

where $\mathcal{U}_L^*$ denotes the set of all counterfactual leak nodes for the symptoms $\mathcal{S}_A^*, \mathcal{S}_B^*, \mathcal{S}_C^*$. First, note that

$$\begin{aligned} &p(S_A^* = 0, S_B^* = 1, S_C^* = 0|\mathcal{S}_\pm, \mathcal{R}, \mathrm{do}(\wedge_{i\neq k} D_i^* = 0), \mathrm{do}(\mathcal{U}_L^* = 0)) \\ &= \frac{p(S_A^* = 0, S_B^* = 1, S_C^* = 0, \mathcal{S}_\pm|\mathcal{R}, \mathrm{do}(\wedge_{i\neq k} D_i^* = 0), \mathrm{do}(\mathcal{U}_L^* = 0))}{p(\mathcal{S}_\pm|\mathcal{R}, \mathrm{do}(\wedge_{i\neq k} D_i^* = 0), \mathrm{do}(\mathcal{U}_L^* = 0))} \\ &= \frac{p(S_A^* = 0, S_B^* = 1, S_C^* = 0, \mathcal{S}_\pm|\mathcal{R}, \mathrm{do}(\wedge_{i\neq k} D_i^* = 0), \mathrm{do}(\mathcal{U}_L^* = 0))}{p(\mathcal{S}_\pm|\mathcal{R})} \end{aligned}$$

Which follows from the fact that the factual symptoms $\mathcal{S}_\pm$ on the twin network [F] are conditionally independent from the counterfactual interventions $\text{do}(\wedge_{i\neq k} D_i^* = 0), \text{do}(\mathcal{U}_L^* = 0))$. To determine $Q = p(S_A^* = 0, S_B^* = 1, S_C^* = 0, \mathcal{S}_\pm | \mathcal{R}, \text{do}(\wedge_{i\neq k} D_i^* = 0), \text{do}(\mathcal{U}_L^* = 0))$, we express Q as a marginalization over the factual diseases which, together with the interventions on the counterfactual diseases and leak nodes, constitute a Markov blanket for each dual pair of symptoms

$$
\begin{aligned}
Q = \sum_{d_1,\ldots,d_N} & p(\wedge_{i\neq k} D_i = d_i, D_k = d_k | \mathcal{R}) \prod_{S\in\mathcal{S}_A} p(S^* = 0, S = 0 | \wedge_{i\neq k} D_i = d_i, D_k = d_k, \text{do}(\wedge_{i\neq k} D_i^* = 0), \text{do}(\mathcal{U}_L^* = 0)) \\
& \times \prod_{S\in\mathcal{S}_B} p(S^* = 1, S = 1 | \wedge_{i\neq k} D_i = d_i, D_k = d_k, \text{do}(\wedge_{i\neq k} D_i^* = 0), \text{do}(\mathcal{U}_L^* = 0)) \\
& \times \prod_{S\in\mathcal{S}_C} p(S^* = 0, S = 1 | \wedge_{i\neq k} D_i = d_i, D_k = d_k, \text{do}(\wedge_{i\neq k} D_i^* = 0), \text{do}(\mathcal{U}_L^* = 0))
\end{aligned} \tag{26}
$$

Substituting in the CPT derived in Lemma 56 yields

$$
Q = \sum_{d_1,\ldots,d_N} p(\wedge_{i\neq k} D_i = d_i, D_k = d_k | \mathcal{R}) \prod_{S\in\mathcal{S}_A} p(s = 0 | \wedge_{i=1}^N d_i) \prod_{S\in\mathcal{S}_B} (1-\lambda_{D_k,S})\delta(d_k - 1) \prod_{S\in\mathcal{S}_C} \lambda_{D_k,s}^{d_k} p(s^{\setminus k} = 1 | \wedge_{i\neq k} d_i, d_k = 1) \tag{27}
$$

The only terms in (24) with $|\mathcal{S}'_+| \neq 0$ have $\mathcal{S}_B \neq \emptyset$, therefore the term $\delta(d_k - 1)$ is present, and Q simplifies to

$$
Q = \sum_{d_i \forall i\neq k} p(\wedge_{i\neq k} D_i = d_i, D_k = 1 | \mathcal{R}) \prod_{S\in\mathcal{S}_A} p(s = 0 | \wedge_{i\neq k}^N d_i, d_k = 1) \prod_{S\in\mathcal{S}_B} (1 - \lambda_{D_k,S}) \prod_{S\in\mathcal{S}_C} \lambda_{D_k,s} p(s^{\setminus k} = 1 | \wedge_{i\neq k} d_i, d_k = 1) \tag{28}
$$

$$
= p(S_A = 0, S_C^{\setminus k} = 1, D_k = 1 | \mathcal{R}) \prod_{s\in S_B} (1 - \lambda_{D_k,S}) \prod_{s\in S_C} \lambda_{D_k,S} \tag{29}
$$

where in the last line we have performed the marginalization over $d_i \ \forall \ i \neq k$. Finally, $\mathcal{S}'_+ = \mathcal{S}_B^* = \mathcal{S}_+ \setminus \mathcal{S}_C$, and so $|\mathcal{S}'_+| = |\mathcal{S}_+| - |\mathcal{S}_C|$, and the expected expected sufficiency is

$$
\mathbb{E}_{\text{suff}}(D_K, \mathcal{E}) = \sum_{\mathcal{S}\subseteq\mathcal{S}_+} (|\mathcal{S}_+| - |\mathcal{S}|) \frac{p(\mathcal{S}_- = 0, \mathcal{S}^{\setminus k} = 1, D_k = 1 | \mathcal{R})}{p(\mathcal{S}_\pm | \mathcal{R})} \prod_{s\in\mathcal{S}_+\setminus\mathcal{S}} (1 - \lambda_{k,s}) \prod_{s\in\mathcal{S}} \lambda_{k,s} \tag{30}
$$

where we have dropped the subscript C from $\mathcal{S}_C$.

□

Given our expression for the expected sufficiency, we now derive a simplified expression that is very similar to the posterior $p(d_k = 1 | \mathcal{R}, \mathcal{S}_\pm)$.

Theorem 4 (Simplified expected sufficiency)**.**

$$
\mathbb{E}_{suff}(D_K, \mathcal{E}) = \frac{1}{p(\mathcal{S}_\pm | \mathcal{R})} \sum_{\mathcal{Z}\subseteq\mathcal{S}_+} (-1)^{|\mathcal{Z}|} p(\mathcal{S}_- = 0, \mathcal{Z} = 0, d_k = 1 | \mathcal{R}) \times \tau(k, \mathcal{Z}) \tag{31}
$$

where

$$
\tau(k, \mathcal{Z}) = \sum_{S\in\mathcal{S}_+\setminus\mathcal{Z}} (1 - \lambda_{D_k,S}) \tag{32}
$$

Proof. Starting with the expected sufficiency given in 2, we can perform the change of variables $\mathcal{X} = \mathcal{S}_+ \setminus \mathcal{S}$ to give

$$
\mathbb{E}_{\text{suff}}(D_K, \mathcal{E}) = \frac{1}{p(\mathcal{S}_\pm | \mathcal{R})} \sum_{\mathcal{X}\subseteq\mathcal{S}_+} |X| \prod_{S\in\mathcal{X}} (1 - \lambda_{D_k,S}) \prod_{S\in\mathcal{S}_+\setminus X} \lambda_{D_k,S}\, p(\mathcal{S}_- = 0, (\mathcal{S}_+ \setminus \mathcal{X})^{\setminus k} = 1, d_k = 1 | \mathcal{R}) \tag{33}
$$

$$
= \frac{1}{p(\mathcal{S}_\pm | \mathcal{R})} \sum_{\mathcal{X}\subseteq\mathcal{S}_+} |\mathcal{X}| \prod_{S\in\mathcal{X}} (1 - \lambda_{D_k,S}) \prod_{S\in\mathcal{S}_+\setminus\mathcal{X}} \lambda_{D_k,S} \sum_{\mathcal{Z}\subseteq\mathcal{S}_+\setminus\mathcal{X}} (-1)^{|\mathcal{Z}|} p(\mathcal{S}_- = 0, \mathcal{Z}^{\setminus k} = 0, d_k = 1 | \mathcal{R}) \tag{34}
$$

where in the last line we apply the inclusion-exclusion principle to decompose an arbitrary joint state over Bernoulli variables $p(\mathcal{A}=0,\mathcal{B}=1)$ as a sum over the powerset of the variables $\mathcal{B}$ in terms of marginals where all variables are instantiated to 0,

$$p(\mathcal{A}=0,\mathcal{B}=1)=\sum_{\mathcal{C}\subseteq\mathcal{B}}(-1)^{|C|}p(\mathcal{A}=0,\mathcal{C}=0) \tag{35}$$

By the definition of noisy-or (11) we have that

$$\begin{aligned}
&p(\mathcal{S}_-=0,\mathcal{Z}^{\setminus k}=0,d_k=1|\mathcal{R})\\
&\quad=\sum_{d_i,i\neq k}p(\mathcal{S}_-=0,\mathcal{Z}^{\setminus k}=0,d_k=1,\wedge_{i\neq k}^N D_i=d_i|\mathcal{R})\\
&\quad=\sum_{d_i,i\neq k}\prod_{S\in\mathcal{S}_-}p(S=0|d_k=1,\wedge_{i\neq k}^N D_i=d_i)\prod_{S\in\mathcal{Z}}p(S^{\setminus k}=0|d_k=1,\wedge_{i\neq k}^N D_i=d_i)p(d_k=1,\wedge_{i\neq k}^N D_i=d_i|\mathcal{R})\\
&\quad=\sum_{d_i,i\neq k}\prod_{S\in\mathcal{S}_-}p(S=0|d_k=1,\wedge_{i\neq k}^N D_i=d_i)\prod_{S\in\mathcal{Z}}\frac{p(S=0|d_k=1,\wedge_{i\neq k}^N D_i=d_i)}{\lambda_{D_k,S}}p(d_k=1,\wedge_{i\neq k}^N D_i=d_i|\mathcal{R})\\
&\quad=\frac{p(\mathcal{S}_-=0,\mathcal{Z}=0,d_k=1|\mathcal{R})}{\prod_{S\in\mathcal{Z}}\lambda_{D_k,S}}
\end{aligned} \tag{36}$$

Therefore we can replace the graph operation represented by $\setminus k$ by dividing the CPT by the product $\prod_{S\in Z}\lambda_{D_k,S}$. Denote $\mathcal{Y}=\mathcal{S}_+\setminus\mathcal{X}$. This allows $\mathbb{E}_{\text{suff}}$ to be expressed as

$$\mathbb{E}_{\text{suff}}(D_K,\mathcal{E})=\frac{1}{p(\mathcal{S}_\pm|\mathcal{R})}\sum_{\mathcal{X}\subseteq\mathcal{S}_+}|\mathcal{X}|\prod_{S\in\mathcal{X}}(1-\lambda_{D_k,S})\prod_{S\in\mathcal{Y}}\lambda_{D_k,S}\sum_{\mathcal{Z}\subseteq\mathcal{Y}}(-1)^{|\mathcal{Z}|}p(\mathcal{S}_-=0,\mathcal{Z}=0,d_k=1|\mathcal{R})\frac{1}{\prod_{S\in\mathcal{Z}}\lambda_{D_k,S}} \tag{37}$$

We now aggregate terms in the power sum that yield the same marginal on the symptoms (e.g. for fixed $\mathcal{Z}$). Every $\mathcal{X}\in\mathcal{S}_+\setminus\mathcal{Z}$ yields a single marginal $p(\mathcal{S}_-=0,\mathcal{Z}=0,d_k=1|\mathcal{R})$ and therefore if we express (37) as a sum in terms of $\mathcal{Z}$, where each term $p(\mathcal{S}_-=0,\mathcal{Z}=0,d_k=1|\mathcal{R})$ aggregates the a coefficient K of the form

$$\begin{aligned}
K&=\frac{(-1)^{|\mathcal{Z}|}}{p(\mathcal{S}_\pm|\mathcal{R})}\frac{1}{\prod_{S\in\mathcal{Z}}\lambda_{D_k,S}}\sum_{\mathcal{X}\subseteq\mathcal{S}_+\setminus\mathcal{Z}}|\mathcal{X}|\prod_{S\in\mathcal{X}}(1-\lambda_{D_k,S})\prod_{S\in S_+\setminus\mathcal{X}}\lambda_{D_k,S}\\
&=\frac{(-1)^{|\mathcal{Z}|}}{p(\mathcal{S}_\pm|\mathcal{R})}\frac{1}{\prod_{S\in\mathcal{Z}}\lambda_{D_k,S}}\sum_{\mathcal{X}\subseteq\mathcal{A}}|\mathcal{X}|\prod_{S\in\mathcal{X}}(1-\lambda_{D_k,S})\prod_{S\in\mathcal{A}\setminus\mathcal{X}}\lambda_{D_k,S}\prod_{S\in\mathcal{Z}}\lambda_{D_k,S}\\
&=\frac{(-1)^{|\mathcal{Z}|}}{p(\mathcal{S}_\pm|\mathcal{R})}\sum_{\mathcal{X}\subseteq\mathcal{A}}|\mathcal{X}|\prod_{S\in\mathcal{X}}(1-\lambda_{D_k,S})\prod_{S\in\mathcal{A}\setminus\mathcal{X}}\lambda_{D_k,S}
\end{aligned} \tag{38}$$

where $\mathcal{A}=\mathcal{S}_+\setminus\mathcal{Z}$. This can be further simplified using the identity

$$\sum_{A\subseteq B}|A|\prod_{a\in A}(1-a)\prod_{a'\in B\setminus A}a'=|B|-\sum_{a\in B}a=\sum_{a\in B}(1-a) \tag{39}$$

which we now prove iteratively. First, consider the function $S(\mathcal{B}):=\sum_{\mathcal{A}\subseteq\mathcal{B}}\prod_{a\in\mathcal{A}}(1-a)\prod_{a'\in\mathcal{B}\setminus\mathcal{A}}a'$. Now, consider $S(\mathcal{B}+\{c\})$. This function can be divided into two sums, one where $c\in\mathcal{A}$ and the other where $c\notin\mathcal{A}$. Therefore

$$S(\mathcal{B}+\{c\})=\sum_{\mathcal{A}\subseteq\mathcal{B}}\prod_{a\in\mathcal{A}}(1-a)\prod_{a'\in\mathcal{B}\setminus\mathcal{A}}a'c+\sum_{\mathcal{A}\subseteq\mathcal{B}}\prod_{a\in\mathcal{A}}(1-a)\prod_{a'\in\mathcal{B}\setminus\mathcal{A}}a'(1-c)=S(\mathcal{B}) \tag{40}$$

Starting with the empty set, $S(\emptyset) = 1$, it follows that $S(\mathcal{B}) = 1\ \forall$ countable sets $\mathcal{B}$. Next, consider the function $G(\mathcal{B}) := \sum_{\mathcal{A}\subseteq\mathcal{B}} |\mathcal{A}| \prod_{a\in\mathcal{A}} (1-a) \prod_{a'\in\mathcal{B}\setminus\mathcal{A}} a'$, which is the form of the sum we wish to compute in (38). Proceeding as before, we have

$$\begin{aligned} G(\mathcal{B}+\{c\}) &= \sum_{\mathcal{A}\subseteq B} |\mathcal{A}| \prod_{a\in\mathcal{A}} (1-a) \prod_{a'\in\mathcal{B}\setminus\mathcal{A}} a'c + \sum_{\mathcal{A}\subseteq\mathcal{B}} (|\mathcal{A}|+1) \prod_{a\in\mathcal{A}} (1-a) \prod_{a'\in\mathcal{B}\setminus\mathcal{A}} a'(1-c) \\ &= cG(\mathcal{B}) + (1-c)G(\mathcal{B}) + (1-c)S(\mathcal{B}) \end{aligned}$$

Using $S(\mathcal{B}) = 1$ we arive at the recursive formula $G(\mathcal{B}+\{c\}) = G(\mathcal{B})+(1-c)$. Starting with $G(\emptyset) = 0$, and building the set $\mathcal{B}$ by recursively adding elements c to the set, we arrive at the identity

$$G(\mathcal{B}) = |\mathcal{B}| - \sum_{a\in\mathcal{B}} a \tag{41}$$

Using (41) we can simplify the coefficient (38)

$$\frac{(-1)^{|\mathcal{Z}|}}{p(\mathcal{S}_\pm|\mathcal{R})} \sum_{\mathcal{X}\subseteq\mathcal{S}_+\setminus\mathcal{Z}} |\mathcal{X}| \prod_{S\in\mathcal{X}} (1-\lambda_{D_k,S}) \prod_{S\in(\mathcal{S}_+\setminus\mathcal{Z})\setminus\mathcal{X}} \lambda_{D_k,S} = \frac{(-1)^{|\mathcal{Z}|}}{p(\mathcal{S}_\pm|\mathcal{R})} \sum_{S\in S_+\setminus\mathcal{Z}} (1-\lambda_{D_k,S}) \tag{42}$$

Rearranging (37) as a summation over $\mathcal{Z}$ substituting in (42) gives

$$\mathbb{E}_{\text{suff}}(D_K,\mathcal{E}) = \frac{1}{p(\mathcal{S}_\pm|\mathcal{R})} \sum_{\mathcal{Z}\subseteq\mathcal{S}_+} (-1)^{|\mathcal{Z}|} p(\mathcal{S}_- = 0, \mathcal{Z} = 0, d_k = 1|\mathcal{R}) \left(\sum_{S\in\mathcal{S}_+\setminus\mathcal{Z}} (1-\lambda_{D_k,S}) \right) \tag{43}$$

which can be expressed as

$$\mathbb{E}_{\text{suff}}(D_K,\mathcal{E}) = \frac{1}{p(\mathcal{S}_\pm|\mathcal{R})} \sum_{\mathcal{Z}\subseteq\mathcal{S}_+} (-1)^{|\mathcal{Z}|} p(\mathcal{S}_- = 0, \mathcal{Z} = 0, d_k = 1|\mathcal{R}) \times \tau(k,\mathcal{Z}) \tag{44}$$

where

$$\tau(k,\mathcal{Z}) = \sum_{S\in\mathcal{S}_+\setminus\mathcal{Z}} (1-\lambda_{D_k,S}) \tag{45}$$

Note that if we fix $\tau(k,\mathcal{Z}) = 1\ \forall\mathcal{Z}$, we recover $\sum_{\mathcal{Z}\subseteq\mathcal{S}_+} (-1)^{|\mathcal{Z}|} p(\mathcal{S}_- = 0, \mathcal{Z} = 0, d_k = 1|\mathcal{R})/p(\mathcal{S}_\pm|\mathcal{R}) = p(\mathcal{S}_\pm, d_k = 1|\mathcal{R})/p(\mathcal{S}_\pm|\mathcal{R}) = p(d_k = 1|\mathcal{E})$, which is the standard posterior of disease D_k under evidence $\mathcal{E} = \mathcal{R}\cap\mathcal{S}_\pm$. Note that (44) can be seen as a counterfactual correction to the quickscore algorithm in [66] (although we do not assume independence of diseases as the authors of [66] do).

□

E. PROPERTIES OF THE EXPECTED SUFFICIENCY

In this appendix, we show that the expected sufficiency (46) obeys our four postulates, including an additional postulate of *sufficiency* which is obeyed by the expected sufficiency.

1. *consistency.* $\mathbb{E}_{\text{suff}}(D_K,\mathcal{E}) \propto p(D_k = 1|\mathcal{E})$

2. *causality.* If $\nexists\ S \in \mathsf{Dec}(D_k)\cap\mathcal{S}_+ \implies E_{\text{suff}}(D_K,\mathcal{E}) = 0$

3. *simplicity.* $|E_{\text{suff}}(D_K, \mathcal{E})| \leq |\mathcal{S}_+ \cap \mathsf{Dec}(D_k)|$

4. *sufficiency.* $E_{\text{suff}}(D_i \wedge D_j, \mathcal{E}) > 0 \implies E_{\text{suff}}(D_i, \mathcal{E}) > 0$ and $E_{\text{suff}}(D_j, \mathcal{E}) > 0$

Postulate 1 dictates that the measure should proportional to the posterior probability of the diseases. Postulate 2 states that if the disease has no causal effect on the symptoms presented then it is a poor diagnosis and should be discarded. Postulate 3 states that the (tight) upper bound of the measure for a given disease (in the sense that there exists some disease model that achieves this upper bound – namely deterministic models) is the number of positive symptoms that the disease can explain. This allows us to differentiate between diseases that are equally likely causes, but where one can explain more symptoms than another. Postulate 4 states that if it is possible that D_k is causing at least one symptom, then the measure should be strictly greater than 0. Starting from the definition of the expected sufficiency

$$\mathbb{E}_{\text{suff}}(D_K, \mathcal{E}) := \sum_{\mathcal{S}'} \left|\mathcal{S}'_+\right| p(\mathcal{S}'|\mathcal{E}, \text{do}(\mathcal{D} \setminus D_k = 0), \text{do}(\mathcal{U}_L = 0)) \tag{46}$$

given the conditional independence structure of the twin network [F], we can express the counterfactual symptom margainls as

$$\begin{aligned} & p(\mathcal{S}'|\mathcal{E},\text{do}(\mathcal{D} \setminus D_k = 0), \text{do}(\mathcal{U}_L = 0)) && (47)\\ &= \sum_{d_k} \prod_{S^* \in \mathcal{S}'} p(S^*|\mathcal{E}, \text{do}(\mathcal{D}^* \setminus D_k = 0), \text{do}(\mathcal{U}^*_L = 0), d_k) p(d_k|\mathcal{E}, \text{do}(\mathcal{D}^* \setminus D_k = 0), \text{do}(\mathcal{U}^*_L = 0)) && (48)\\ &= \sum_{d_k} \prod_{S \in \mathcal{S}'} p(S^*|\mathcal{E}, \text{do}(\mathcal{D}^* \setminus D_k = 0), \text{do}(\mathcal{U}^*_L = 0), d_k) p(d_k|\mathcal{E}) && (49) \end{aligned}$$

If $D_k = 1$, then the counterfactual states have all parents (including leaks) instantiated to 0, which implies that $\mathcal{S}'_= = \emptyset$ by (4). For $D_k = 1$, we recover that $p(\mathcal{S}'|\mathcal{E}, \text{do}(\mathcal{D} \setminus D_k = 0), \text{do}(\mathcal{U}_L = 0)) \propto p(D_k = 1|\mathcal{E})$ and therefore $\mathbb{E}_{\text{suff}}(D_K, \mathcal{E}) \propto p(D_k = 1|\mathcal{E})$, which proves the correct behaviour in the limit $p(D_k = 1|\mathcal{E}) \to 0$ to be consistent with postulate 1. For postulate 2, if there are no symptoms that are descendants of D_k, then $\mathbb{E}_{\text{suff}}(D_K, \mathcal{E}) = 0$. This follows immediately from the fact that if D_k is not an ancestor of any of the symptoms, then all counterfactual symptoms have all parents instantiated as 0 and $\mathcal{S}'_+ = \emptyset$. For postulate 4, we can only prove this property under additional assumptions about our disease model (see appendix B for noisy-and counter example). First, note that $E_{\text{suff}}(D_K, \mathcal{E})$ is a convex sum with positive semi-definite coefficients $|\mathcal{S}'_+|$. If there is a single positively evidenced symptom that is a descendent of D_k, and D_k has a positive causal influence on that child, and our disease model permits that every disease be capable of causing its associated symptoms in isolation, i.e. $p(S = 1|\text{only})(D_k = 1)) > 0$ for $S \in \mathsf{Dec}(D_k)$, then it is simple to check that $p(S^* = 1|\mathcal{E}, \text{do}(\mathcal{D}^* \setminus D_k = 0), \text{do}(\mathcal{U}^*_L = 0), d_k = 1) > 0$ and so $E_{\text{suff}}(D_K, \mathcal{E}) > 0$.

F. EXPECTED DISABLEMENT

In this appendix we turn our attention to our second diagnostic measure – the expected disablement. This measure is closer to typical treatment measures, such as the efffect of treatment on the treated [46]. We use our twin diagnostic network outlined in appendix C figure [E] (shown below) to simulating counterfactual treatments. We focus on the simplest case of single disease interventions, and propose a simple ranking measure whereby the best treatments are those that get rid of the most symptoms.

Definition 1. *The expected disablement of disease D_k determines the number of positively evidenced symptoms that we would expect to switch off if we intervened to switch off D_k*

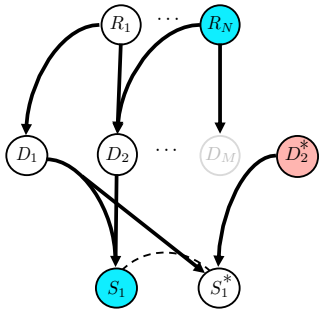
$$\mathbb{E}_{dis}(D_K, \mathcal{E}) := \sum_{\mathcal{S}'_\pm} \left|\mathcal{S}_+ \setminus \mathcal{S}'_+\right| p(\mathcal{S}'_\pm|\mathcal{E}, do(D_k = 0)) \tag{50}$$

where $\mathcal{E}$ is the factual evidence, $\mathcal{S}_+$ is the set of factual positively evidenced diseases and $\mathcal{S}'_+$ is the set of counterfactual symptoms that remain on following the intervention $do(D_k = 0)$.

Decisions about which treatment to select for a patient generally take into account variables such as cost and cruelty. These variables can be simply included in the treatment measure. For example, the cruelty of specific symptoms can be included in the expectation (50) by weighting each positive symptom accordingly. The cost of treating a specific disease is included simply by multiplying (50) by a cost weight, and likewise for including the probability of the intervention succeeding. For now, we focus on computing the counterfactual probabilities, which we can then use to construct arbitrarily weighted expectations.

To calculate (50), note that the only CPTs that differ from the original noisy-OR SCM are those for unmerged dual symptom nodes (i.e. children of the intervention node D_K). The disease layer forms a Markov blanket for the symptoms layer, d-separating dual symptom pairs from each other. Therefore we derive the cpt for dual symptoms and their parent diseases.

[E]



Lemma 2. *For a given symptom S and its counterfactual dual S^*, with parent diseases $\mathcal{D}$ and under the counterfactual intervention $do(D_k^* = 0)$, the joint conditional distribution on the twin network is given by*

$$p(s, s^* \mid \wedge_{i\neq k} D_i = d_i, D_k = d_k, do(D_k^* = 0)) = \begin{cases} p(s = 0 \mid \wedge_i D_i = d_i) & \text{if } s = s^* = 0 \\ 0 \quad \text{if } s = 0, s^* = 1 \\ \left(\frac{1}{\lambda_{D_k,S}} - 1\right) p(s = 0 \mid \wedge_{i\neq k} D_i = d_i, D_k = 1)\delta(d_k - 1) & \text{if } s = 1, s^* = 0 \text{ and } \lambda_{D_k,} \\ p(s^{\setminus k} = 0 \mid \wedge_{i\neq k} D_i = d_i, D_k = 1)\delta(d_k - 1) & \text{if } s = 1, s^* = 0 \text{ and } \lambda_{D_k,S} = 0 \\ p(s^{\setminus k} = 1 \mid \wedge_{i\neq k} D_i = d_i, D_k = 1) & \text{if } s = 1, s^* = 1 \end{cases}$$

where $\delta(d_k - 1) = 1$ if $D_k = 1$ else 0.

Proof. First note that for this marginal distribution the intervention $\text{do}(D_k^* = 0)$ is equivalent to setting the evidence $D_k^* = 0$ as we specify the full Markov blanket of (s, s^*). Let $\mathcal{D}_{\setminus k}$ denote the set of parents of (s, s^*) not including the intervention node D_k^* or its dual D_k. We wish to compute the conditional probability

$$p(s, s^* \mid \wedge_{i\neq k} D_i = d_i, D_k = d_k) = \sum_{u_s} \mathbb{P}(u_s) p(s \mid \wedge_{i\neq k} D_i = d_i, D_k = d_k, u_s) p(s^* \mid \wedge_{i\neq k} D_i = d_i, D_k^* = 0, u_s) \tag{51}$$

We proceed as before by expressing this as a marginalization over the CPT of the dual states, $p(s = 0, s^* = 0 \mid \wedge_{i\neq k} D_i = d_i, D_k^* = 0, D_k)$, $p(s = 0 \mid \wedge_{i\neq k} D_i = d_i, D_k^* = 0, D_k = d_k)$ and $p(s^* = 0 \mid \wedge_{i\neq k} D_i = d_i, D_k^* = 0, D_k = d_k)$. For $s_i = 0$, the generative functions are given by

$$p(s = 0 \mid \text{Pa(S)}, u_s) = u_L \wedge \bigwedge_{D_i \in \text{Pa(S)}} (\bar{d}_i \vee u_{D_i,S}) \tag{52}$$

First we compute the joint state.

$$\begin{aligned} & p(s = 0 \mid \wedge_{i\neq k} D_i = d_i, D_k = d_k, u_s) p(s^* = 0 \mid \wedge_{i\neq k} D_i = d_i, D_k^* = d_k^*, u_s) \\ & = u_L \wedge u_L \bigwedge_{D_i \in \mathcal{D}_{\setminus k}} (u_{D_i,S} \vee \bar{d}_i) \bigwedge_{D_j \in \mathcal{D}_{\setminus k}} (u_{D_j,S} \vee \bar{d}_j) \wedge \left[u_{D_k,S} \vee \bar{d}_k\right] \wedge \left[u_{D_k,S} \vee \bar{d}_k^*\right] \\ & = u_L \bigwedge_{D_i \in \mathcal{D}_{\setminus k}} (u_{D_i,S} \vee \bar{d}_i) \left[u_{D_k,S} \vee (\bar{d}_k^* \wedge \bar{d}_k)\right] \end{aligned}$$

Where we have used the Boolean identities $a \wedge a = a$ and $a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c)$. Therefore

$$\begin{aligned} p(s=0, s^*=0|\wedge_{i\neq k} D_i = d_i, D_k = d_k, D_k^* = d_k^*) &= \sum_{u_s} \mathbb{P}(u_s) p(s=0|\mathcal{D}_{\setminus k}, D_k, u_s) p(s^*=0|\mathcal{D}_{\setminus k}, D_k^*, u_s) \\ &= \lambda_{L_s}\left[\lambda_{D_k,S}\left(d_k \vee d_k^*\right) + \bar{d}_k \wedge \bar{d}_k^*\right] \prod_{D_i \in \mathcal{D}_{\setminus k}} \left[\lambda_{D_i,S} d_i + \bar{d}_i\right] \end{aligned}$$

Next, we calculate the single-symptom conditionals

$$\begin{aligned} p(s=0\,|\,\wedge_{i\neq k} D_i = d_i, D_k = d_k) &= \sum_{u_s} \mathbb{P}(u_s) p(s=0|\mathcal{D}_{\setminus k}, D_k, u_s) \\ &= \sum_{u_{L_S}} p(u_{L_s}) u_{L_s} \prod_{D_i \in \mathcal{D}} \sum_{u_{D_i,S}} p(u_{D_i,S}) u_{D_i,S} \vee \bar{d}_i \\ &= p(u_{L_s}=1) \prod_{D_i \in \mathcal{D}} \sum_{u_{D_i,S}} \left[p(u_{D_i,S}=1) + p(u_{D_i,S}=0)\bar{d}_i\right] \\ &= \lambda_{L_s} \prod_{D_i \in \mathcal{D}} \left[\lambda_{D_i,S} d_i + \bar{d}_i\right] \end{aligned} \tag{53}$$

and similar for $p(s^*=0\,|\,\wedge_{i\neq k} D_i = d_i, D_k^* = d_k^*)$. Note that $\lambda x + \bar{x} = \lambda^x$. We can now express the joint cpd over dual symptom pairs, using the identities $p(s=0, s^*=1\,|\,X) = p(s=0\,|\,X) - p(s=0, s^*=0\,|\,X)$, $p(s=1, s^*=0\,|\,X) = p(s^*=0\,|\,X) - p(s=0, s^*=0\,|\,X)$ and $p(s=1, s^*=1\,|\,X) = 1 - p(s=0\,|\,X) - p(s^*=0\,|\,X) + p(s=0, s^*=0\,|\,X)$ for arbitrary conditional X.

$$p(s, s^*|\wedge_{i\neq k} D_i = d_i, D_k = d_k, D_k^* = d_k^*) = \begin{cases} \lambda_{L_s} \lambda_{D_k,S}^{d_k \vee d_k^*} \prod_{D_i \in \mathcal{D}_{\setminus k}} \lambda_{D_i,S}^{d_i} & \text{if } s = s^* = 0 \\ \lambda_{L_s}\left[\lambda_{D_k,S}^{d_k} - \lambda_{D_k,S}^{d_k \vee d_k^*}\right] \prod_{D_i \in \mathcal{D}_{\setminus k}} \lambda_{D_i,S}^{d_i} & \text{if } s = 0, s^* = 1 \\ \lambda_{L_s}\left[\lambda_{D_k,S}^{d_k^*} - \lambda_{D_k,S}^{d_k \vee d_k^*}\right] \prod_{D_i \in \mathcal{D}_{\setminus k}} \lambda_{D_i,S}^{d_i} & \text{if } s = 1, s^* = 0 \\ 1 - \lambda_{L_s}\left[\lambda_{D_k,S}^{d_k} + \lambda_{D_k,S}^{d_k^*} - \lambda_{D_k,S}^{d_k \vee d_k^*}\right] \prod_{D_i \in \mathcal{D}_{\setminus k}} \lambda_{D_i,S}^{d_i} & \text{if } s = s^* = 1 \end{cases}$$

As we are always intervening to switch off diseases, $D_k^* = 0$, then $d_k \vee d_k^* = d_k$ and

$$\lambda_{D_k,S}^{d_k} - \lambda_{D_k,S}^{d_k \vee d_k^*} = 0 \tag{54}$$

and therefore $p(s=0, s^*=1|\wedge_{i\neq k} D_i = d_i, D_k = d_k, D_k^* = 0) = 0$ as expected (switching off a disease will never switch on a symptom). This simplifies our expression for the conditional distribution to

$$p(s, s^*|\wedge_{i\neq k} D_i = d_i, D_k = d_k, D_k^* = 0) = \begin{cases} \lambda_{L_s} \lambda_{D_k,S}^{d_k} \prod_{D_i \in \mathcal{D}_{\setminus k}} \lambda_{D_i,S}^{d_i} & \text{if } s = s^* = 0 \\ 0 \quad \text{if } s = 0, s^* = 1 \\ \lambda_{L_s}\left[1 - \lambda_{D_k,S}^{d_k}\right] \prod_{D_i \in \mathcal{D}_{\setminus k}} \lambda_{D_i,S}^{d_i} & \text{if } s = 1, s^* = 0 \\ 1 - \lambda_{L_s} \prod_{D_i \in \mathcal{D}_{\setminus k}} \lambda_{D_i,S}^{d_i} & \text{if } s = s^* = 1 \end{cases} \tag{55}$$

This then simplifies using (53) to

$$p(s, s^*|\wedge_{i\neq k} D_i = d_i, D_k = d_k, D_k^* = 0) = \begin{cases} p(s=0|\wedge_i D_i = d_i) & \text{if } s = s^* = 0 \\ 0 \quad \text{if } s = 0, s^* = 1 \\ p(s=0|\wedge_{i\neq k} D_i = d_i, D_k = 0) - p(s=0|\wedge_{i\neq k} D_i = d_i, D_k = d_k) & \text{if } s = 1, s^* = 0 \\ p(s=1|\wedge_{i\neq k} D_i = d_i, D_k = 0) & \text{if } s = s^* = 1 \end{cases} \tag{56}$$

We have arrived at expressions for the cpt's over dual symptoms in terms of cpt's on the factual graph, and hence our conterfactual query can be computed on the factual graph alone. The third term in (56), $p(s=0|\wedge_{i\neq k} D_i = d_i, D_k = 0) - p(s=0|\wedge_{i\neq k} D_i = d_i D_k = d_k)$, equals zero unless $d_k = 1$. Using the definition of noisy-OR (11) to give

$$p(s=0|\wedge_{i\neq k} D_i = d_i, D_k = 0) = \frac{1}{\lambda_{D_k,S}} p(s=0|\wedge_{i\neq k} D_i = d_i, D_k = 1) \tag{57}$$

in the case that $\lambda_{D_k,S} > 0$, we recover

$$p(s=0|\wedge_{i\neq k} D_i = d_i, D_k = 0) - p(s=0|\wedge_{i\neq k} D_i = d_i, D_k = d_k) = \left(\frac{1}{\lambda_{D_k,S}} - 1\right) p(s=0\,|\wedge_{i\neq k} D_i = d_i, D_k = 1)\delta(d_k - 1) \tag{58}$$

where d_k is the instantiation of D_k on the factual graph. The term $\delta(d_k - 1)$ is equivalent to fixing the observation $D_k = 1$ on the factual graph. If $\lambda_{D_k,S} = 0$ then

$$\lambda_{L_s}\left[1 - \lambda_{D_k,S}^{d_k}\right] \prod_{D_i \in \mathcal{D}_{\setminus k}} \lambda_{D_i,S}^{d_i} = \lambda_{L_s} \prod_{D_i \in \mathcal{D}_{\setminus k}} \lambda_{D_i,S}^{d_i} \delta(d_k - 1) \tag{59}$$

which is equivalent to $p(s^{\setminus k} = 0|\wedge_{i\neq k} D_i = d_i, D_k = 1)\delta(d_k - 1)$
Finally, from the definition of the noisy-OR CPT (4),

$$p(s=1|\wedge_{i\neq k} D_i = d_i, D_k = 0) = p(s^{\setminus k} = 1|\wedge_{i\neq k} D_i = d_i, D_k = 1) \tag{60}$$

□

Lemma 56 allows us to express the expected disablement in terms of factual probabilities. As we have seen, the intervention $\mathrm{do}(D_k^* = 0)$ can never result in counterfactual symptoms that are on, when their dual factual symptoms are off, so we need only enumerate over counterfactual symptoms states where $\mathcal{S}'_+ \subseteq \mathcal{S}_+$ as these are the only counterfactual states with non-zero weight. From this it also follows that for all $s \in \mathcal{S}_- \implies s^* \in \mathcal{S}'_-$. The counterfactual CPT in (50) is represented on the twin network [F] as

$$p(\mathcal{S}'_+, \mathcal{S}'_-|\mathcal{E}, \mathrm{do}(D_k^* = 0)) = p(\mathcal{S}'_+, \mathcal{S}'_-|\mathcal{S}_+, \mathcal{S}_-, \mathcal{R}, \mathrm{do}(D_k^* = 0)) \tag{61}$$

Theorem 5 (Simplified noisy-OR expected disablement)**.** *For the noisy-OR networks described in Appendix B, the expected disablement of disease D_k is given by*

$$\mathbb{E}(D_K, \mathcal{E})_{dis} = \frac{1}{p(\mathcal{S}_+, \mathcal{S}_-|\mathcal{R})} \sum_{\mathcal{Z}\subseteq\mathcal{S}_+} (-1)^{|\mathcal{Z}|} p(\mathcal{S}_- = 0, \mathcal{Z} = 0, D_k = 1|\mathcal{R})\gamma(\mathcal{Z}, D_k) \tag{62}$$

where

$$\gamma(\mathcal{Z}, D_k) = \sum_{S\in\mathcal{Z}} \left(1 - \frac{1}{\lambda_{D_k,S}}\right) \tag{63}$$

where $\mathcal{S}_\pm$ is the set of factual positive (negative) evidenced symptom nodes and $\mathcal{R}$ is the risk factor evidence.

Proof. From the above discussion, the non-zero contributions to the expected disablement are

$$\mathbb{E}(D_K, \mathcal{E})_{\mathrm{dis}} = \sum_{\mathcal{C}\subseteq\mathcal{S}_+} |\mathcal{C}| p(\mathcal{S}_-^* = 0, \mathcal{C}^* = 0, \mathcal{S}_+ \setminus \mathcal{C} = 1|\mathcal{S}_+, \mathcal{S}_-, \mathcal{R}, \mathrm{do}(D_k^* = 0)) \tag{64}$$

Applying Bayes rule, and noting the the factual evidence states are not children of the intervention node D_K^*, gives

$$\mathbb{E}(D_K,\mathcal{E})_{\text{dis}} = \frac{1}{p(\mathcal{S}_+,\mathcal{S}_-|\mathcal{R})}\sum_{\mathcal{C}\subseteq\mathcal{S}_+}|\mathcal{C}|p(\mathcal{S}_-^*=0,\mathcal{C}^*=0,\mathcal{S}_+\setminus\mathcal{C}=1,\mathcal{S}_+,\mathcal{S}_-|\mathcal{R},\text{do}(D_k^*=0)) \tag{65}$$

Let us now consider the probabilities $Q = p(\mathcal{S}_-^*=0,\mathcal{C}^*=0,\mathcal{S}\setminus\mathcal{C}^*=1,\mathcal{S}_+,\mathcal{S}_-|\mathcal{R},\text{do}(D_k^*=0))$. We can express these as marginalizations over the disease layer, which d-separate dual symptom pairs from each-other. First, we express Q in the instance where we assume all $\lambda_{D_k,S}>0$.

$$\begin{aligned}Q = &\sum_{d,d_k} p(\wedge_{i\neq k}D_i=d_i,D_k=d_k|\mathcal{R})\prod_{S\in\mathcal{S}_-}p(S^*=0,S=0|\wedge_{i\neq k}D_i=d_i,D_k=d_k,D_k^*=0)\\
&\times\prod_{S\in\mathcal{C}}p(S^*=0,S=1|\wedge_{i\neq k}D_i=d_i,D_k=d_k,D_k^*=0)\prod_{S\in\mathcal{S}_+\setminus\mathcal{C}}p(S^*=1,S=1|\wedge_{i\neq k}D_i=d_i,D_k=d_k,D_k^*=0)\end{aligned} \tag{66}$$

$\mathbb{E}(D_K,\mathcal{E})$ is a sum of the product terms in Q, therefore if all for these terms are continuous for $\lambda_{D_k,S}\to 0\ \forall\ S$ we can derive $\mathbb{E}(D_K,\mathcal{E})$ for positive $\lambda_{D_k,S}$ and take the limit $\lambda_{D_k,S}\to 0$ where appropriate. We can consider each term in isolation, as the product of continuous functions is continuous. Each term in Q derives from one of

$$\begin{aligned}&p(s,s^*\mid\wedge_{i\neq k}D_i=d_i,D_k=d_k,\text{do}(D_k^*=0))\\
&\quad=\begin{cases}p(s=0\mid\wedge_i D_i=d_i) & \text{if } s=s^*=0\\ 0 & \text{if } s=0,s^*=1\\ \left(\frac{1}{\lambda_{D_k,S}}-1\right)p(s=0\mid\wedge_{i\neq k}D_i=d_i,D_k=1)\delta(d_k-1) & \text{if } s=1,s^*=0 \text{ and } \lambda_{D_k,S}\neq 0\\ p(s^{\setminus k}=0\mid\wedge_{i\neq k}D_i=d_i,D_k=1)\delta(d_k-1) & \text{if } s=1,s^*=0 \text{ and } \lambda_{D_k,S}=0\\ p(s^{\setminus k}=1|\wedge_{i\neq k}D_i=d_i,D_k=1) & \text{if } s=1,s^*=1\end{cases}\end{aligned} \tag{67}$$

Starting with $p(s=0\mid\wedge_i D_i=d_i)=\lambda_{L_S}\prod_{i=1}^N\lambda_{D_i,S}^{d_i}$, this is a linear function of $\lambda_{D_k,S}$ and therefore continuous in the limit $\lambda_{D_k,S}\to 0$. Secondly,

$$\left(\frac{1}{\lambda_{D_k,S}}-1\right)p(s=0\mid\wedge_{i\neq k}D_i=d_i,D_k=1)\delta(d_k-1)=\left(\frac{1}{\lambda_{D_k,S}}-1\right)\lambda_{L_S}\prod_{i=1}^N\lambda_{D_i,S}^{d_i}\delta(d_k-1) \tag{68}$$

which again is a linear function fo $\lambda_{D_k,S}$ and so is continuous in the limit $\lambda_{D_k,S}\to 0$. $p(s^{\setminus k}=0\mid\wedge_{i\neq k}D_i=d_i,D_k=1)\delta(d_k-1)$ is a constant function w.r.t $\lambda_{D_k,S}$, as is $p(s^{\setminus k}=1|\wedge_{i\neq k}D_i=d_i,D_k=1)$, so these are also both continuous in the limit.

We therefore proceed under the assumption that $\lambda_{D_k,S}>0\ \forall\ S$. Applying Lemma 1 simplifies (66) to

$$\begin{aligned}Q = &\sum_d p(\wedge_{i\neq k}D_i=d_i,D_k=d_k|\mathcal{R})\prod_{S\in\mathcal{S}_-}p(S=0|\wedge_{i\neq k}D_i=d_i,D_k=d_k)\prod_{S\in\mathcal{C}}p(S=0|\wedge_{i\neq k}D_i=d_i,D_k=1)\delta(d_k-1)\\
&\times\prod_{S\in\mathcal{S}_+\setminus\mathcal{C}}p(S^{\setminus k}=1|\wedge_{i\neq k}D_i=d_i,D_k=1)\prod_{S\in\mathcal{C}}\left(\frac{1}{\lambda_{D_k,S}}-1\right)\end{aligned} \tag{69}$$

Note that the only Q that are not multiplied by a factor $|\mathcal{C}|=0$ in (65) have $\mathcal{C}\neq\emptyset$, and so $\delta(d_k-1)$ is always present. Marginalizing over all disease states gives

$$Q = p(\mathcal{S}_-=0,\mathcal{C}=0,(\mathcal{S}_+\setminus\mathcal{C})^{\setminus k}=1,D_k=1|\mathcal{R})\prod_{S\in\mathcal{C}}\left(\frac{1}{\lambda_{D_k,S}}-1\right) \tag{70}$$

As before, we simplify this using a change of varaibles and the inclusion-exclusion principle. Change variables $\mathcal{C}\to\mathcal{S}_+\setminus\mathcal{C}$, which along with (70) gives

$$\mathbb{E}(D_K,\mathcal{E})_{\text{dis}} = \frac{1}{p(\mathcal{S}_+,\mathcal{S}_-|\mathcal{R})} \sum_{\mathcal{C}\subseteq\mathcal{S}_+} |\mathcal{S}_+ \setminus \mathcal{C}| p(\mathcal{S}_- = 0, (\mathcal{S}_+ \setminus \mathcal{C}) = 0, \mathcal{C}^{\setminus k} = 1, D_k = 1|\mathcal{R}) \prod_{S\in(\mathcal{S}_+\setminus\mathcal{C})} \left(\frac{1}{\lambda_{D_k,S}} - 1\right) \quad (71)$$

Next we apply the inclusion exclusion principle, giving

$$\mathbb{E}(D_K,\mathcal{E})_{\text{dis}} = \frac{1}{p(\mathcal{S}_+,\mathcal{S}_-|\mathcal{R})} \sum_{\mathcal{C}\subseteq\mathcal{S}_+} |\mathcal{S}_+\backslash\mathcal{C}| \prod_{S\in(\mathcal{S}_+\setminus\mathcal{C})} \left(\frac{1}{\lambda_{D_k,S}} - 1\right) \sum_{\mathcal{Z}\subseteq\mathcal{C}} (-1)^{|\mathcal{Z}|} p(\mathcal{S}_- = 0, (\mathcal{S}_+\backslash\mathcal{C}) = 0, \mathcal{Z}^{\setminus k} = 0, D_k = 1|\mathcal{R}) \quad (72)$$

We can now proceed as before and remove the graph cut operation on the set $\mathcal{Z}$, using the definition of noisy-or (4),

$$\begin{aligned}
&p(\mathcal{S}_- = 0,(\mathcal{S}_+ \setminus \mathcal{C}) = 0, \mathcal{Z}^{\setminus k} = 0, d_k = 1|\mathcal{R}) \\
&\quad = \sum_{d_i, i\neq k} p(\mathcal{S}_- = 0, (\mathcal{S}_+ \setminus \mathcal{C}) = 0, \mathcal{Z}^{\setminus k} = 0, d_k = 1, \wedge_{i\neq k}^N D_i = d_i|\mathcal{R}) \\
&\quad = \sum_{d_i, i\neq k} \prod_{S\in\mathcal{S}_\pm\setminus\mathcal{C}} p(S = 0|d_k = 1, \wedge_{i\neq k}^N D_i = d_i) \prod_{S\in\mathcal{Z}} p(S^{\setminus k} = 0|d_k = 1, \wedge_{i\neq k}^N D_i = d_i) p(d_k = 1, \wedge_{i\neq k}^N D_i = d_i|\mathcal{R}) \\
&\quad = \sum_{d_i, i\neq k} \prod_{S\in\mathcal{S}_\pm\setminus\mathcal{C}} p(S = 0|d_k = 1, \wedge_{i\neq k}^N D_i = d_i) \prod_{S\in\mathcal{Z}} \frac{p(S = 0|d_k = 1, \wedge_{i\neq k}^N D_i = d_i)}{\lambda_{D_k,S}} p(d_k = 1, \wedge_{i\neq k}^N D_i = d_i|\mathcal{R}) \\
&\quad = \frac{p(\mathcal{S}_- = 0, (\mathcal{S}_+ \setminus \mathcal{C}) = 0, \mathcal{Z} = 0, d_k = 1|\mathcal{R})}{\prod_{S\in\mathcal{Z}} \lambda_{D_k,S}}
\end{aligned} \quad (73)$$

Therefore

$$\begin{aligned}
\mathbb{E}(D_K,\mathcal{E})_{\text{dis}} = \frac{1}{p(\mathcal{S}_+,\mathcal{S}_-|\mathcal{R})} \sum_{\mathcal{C}\subseteq\mathcal{S}_+} |\mathcal{S}_+ \setminus \mathcal{C}| \prod_{S\in\mathcal{S}_+\setminus\mathcal{C}} \left(\frac{1}{\lambda_{D_k,S}} - 1\right) \\
\times \sum_{\mathcal{Z}\subseteq\mathcal{C}} (-1)^{|\mathcal{Z}|} p(\mathcal{S}_- = 0, \mathcal{S}_+ \setminus \mathcal{C} = 0, \mathcal{Z} = 0, D_k = 1|\mathcal{R}) \prod_{S\in\mathcal{Z}} \frac{1}{\lambda_{D_k,S}}
\end{aligned}$$

Therefore

$$\mathbb{E}(D_K,\mathcal{E})_{\text{dis}} = \frac{1}{p(\mathcal{S}_+,\mathcal{S}_-|\mathcal{R})} \sum_{\mathcal{C}\subseteq\mathcal{S}_+} |\mathcal{S}_+\backslash\mathcal{C}| \prod_{S\in\mathcal{S}_+\setminus\mathcal{C}} \left(\frac{1}{\lambda_{D_k,S}} - 1\right) \sum_{\mathcal{Z}\subseteq\mathcal{C}} (-1)^{|\mathcal{Z}|} p(\mathcal{S}_- = 0, \mathcal{S}_+\backslash\mathcal{C} = 0, \mathcal{Z} = 0, D_k = 1|\mathcal{R}) \prod_{S\in\mathcal{Z}} \frac{1}{\lambda_{D_k,S}} \quad (74)$$

Finally, we aggregate all terms that have the same symptom marginal. Perform the change of variables $\mathcal{X} = \mathcal{S}_+ \setminus \mathcal{C}$

$$\mathbb{E}(D_K,\mathcal{E})_{\text{dis}} = \frac{1}{p(\mathcal{S}_+,\mathcal{S}_-|\mathcal{R})} \sum_{\mathcal{X}\subseteq\mathcal{S}_+} |\mathcal{X}| \prod_{S\in\mathcal{X}} \left(\frac{1}{\lambda_{D_k,S}} - 1\right) \sum_{\mathcal{Z}\subseteq\mathcal{S}_+\setminus\mathcal{X}} (-1)^{|\mathcal{Z}|} p(\mathcal{S}_- = 0, \mathcal{X} = 0, \mathcal{Z} = 0, D_k = 1|\mathcal{R}) \prod_{S\in\mathcal{Z}} \frac{1}{\lambda_{D_k,S}} \quad (75)$$

Clearly each term for a given $\mathcal{X}$ is zero unless $\lambda_{D_k,S} < 1 \; \forall \, S \in \mathcal{X}$, and so we can restrict ourselves to $\mathcal{S} \subseteq \mathcal{S}_+ \cap \mathsf{Ch}(D_k)$. Furthermore, if any $\lambda_{D_k,S} = 0$ for $S \in \mathcal{X}$, then the symptom marginal (which is linearly dependent on $\lambda_{D_k,S}$) is 0 (there is zero probability of observing this symptom to be off if $D_k = 1$), and this term in the sum is zero. Therefore we can restrict the sum to $\mathcal{X} \subseteq S_+^{(k)}(\lambda > 0)$, where $S_+^{(k)}(\lambda > 0)$ is the set of positively evidenced factual symptoms that are children of D_k and have $\lambda_{D_k,S} > 1$ Let $\mathcal{A} = \mathcal{X} \cup \mathcal{Z}$. Each marginal $p(\mathcal{S}_- = 0, \mathcal{A} = 0, D_k = 1|\mathcal{R})$ aggregates a coefficient

$$\frac{1}{p(\mathcal{S}_+,\mathcal{S}_-|\mathcal{R})}\sum_{\mathcal{X}\subseteq\mathcal{A}}|\mathcal{X}|\prod_{S\in\mathcal{X}}\left(\frac{1}{\lambda_{D_k,S}}-1\right)(-1)^{|\mathcal{A}|-|\mathcal{X}|}\prod_{S\in\mathcal{A}\setminus\mathcal{X}}\frac{1}{\lambda_{D_k,S}} \tag{76}$$

which simplifies to

$$\frac{1}{p(\mathcal{S}_+,\mathcal{S}_-|\mathcal{R})\prod_{S\in\mathcal{A}}\lambda_{D_k,S}}\sum_{\mathcal{X}\subseteq\mathcal{A}}|\mathcal{X}|(-1)^{|\mathcal{A}|-|\mathcal{X}|}\prod_{S\in\mathcal{X}}(1-\lambda_{D_k,S}) \tag{77}$$

To evaluate this term, define the function

$$G(\mathcal{A}):=\sum_{\mathcal{X}\subseteq\mathcal{A}}|\mathcal{X}|(-1)^{|\mathcal{A}|-|\mathcal{X}|}\prod_{S\in\mathcal{X}}(1-\lambda_{D_k,S}) \tag{78}$$

If we append an element $\{\tilde{S}\}$ to the set $\mathcal{A}$, where $\tilde{S}\notin\mathcal{A}$, we can express $G(\mathcal{A}\cup\{\tilde{S}\})$ as

$$G(\mathcal{A}\cup\{\tilde{S}\})=\sum_{\mathcal{X}\subseteq\mathcal{A}}|\mathcal{X}|(-1)^{|\mathcal{A}|+1-|\mathcal{X}|}\prod_{S\in\mathcal{X}}(1-\lambda_{D_k,S})+\sum_{\mathcal{X}\subseteq\mathcal{A}}(|\mathcal{X}|+1)(-1)^{|\mathcal{A}|+1-|\mathcal{X}|-1}\prod_{S\in\mathcal{X}}(1-\lambda_{D_k,S})\,(1-\lambda_{D_k,\tilde{S}}) \tag{79}$$

where we have split the sum into subsets where containing c and not containing c, and then expressed these in terms of the subsets $\mathcal{X}$ of $\mathcal{A}$. This yields the recursive formula

$$G(\mathcal{A}\cup\{c\})=-\lambda_{D_k,\tilde{S}}G(\mathcal{A})+(1-\lambda_{D_k,\tilde{S}})H(\mathcal{A}) \tag{80}$$

where

$$H(\mathcal{A})=\sum_{\mathcal{X}\subseteq\mathcal{A}}(-1)^{|\mathcal{A}|-|\mathcal{X}|}\prod_{S\in\mathcal{X}}(1-\lambda_{D_k,S}) \tag{81}$$

We can determine $H(\mathcal{A})$ by the same technique – noting that

$$\begin{aligned}H(\mathcal{A}\cup\{\tilde{S}\})&=\sum_{\mathcal{X}\subseteq\mathcal{A}}(-1)^{|\mathcal{A}|+1-|\mathcal{X}|}\prod_{S\in\mathcal{X}}(1-\lambda_{D_k,S})+\sum_{\mathcal{X}\subseteq\mathcal{A}}(-1)^{|\mathcal{A}|+1-|\mathcal{X}|-1}\prod_{S\in\mathcal{X}}(1-\lambda_{D_k,S})\,(1-\lambda_{D_k,\tilde{S}})\\&=-H(\mathcal{A})+(1-\lambda_{D_k,\tilde{S}})H(\mathcal{A})\\&=-\lambda_{D_k,\tilde{S}}H(\mathcal{A})\end{aligned}$$

for $\tilde{S}\notin\mathcal{A}$. Then, noting that $H(\emptyset)=1$, we recover

$$H(\mathcal{A})=(-1)^{|\mathcal{A}|}\prod_{S\in\mathcal{A}}\lambda_{D_k,S} \tag{82}$$

and therefore

$$\begin{aligned}G(\mathcal{A}\cup\{\tilde{S}\})&=-\lambda_{D_k,\tilde{S}}G(\mathcal{A})+(1-\lambda_{D_k,\tilde{S}})(-1)^{|\mathcal{A}|}\prod_{S\in\mathcal{A}}\lambda_{D_k,S}\\&=(-1)\left[\lambda_{D_k,\tilde{S}}G(\mathcal{A})+(1-\lambda_{D_k,\tilde{S}})(-1)^{|\mathcal{A}\cup\{\tilde{S}\}|}\prod_{S\in\mathcal{A}}\lambda_{D_k,S}\right]\end{aligned}$$

The above recursion relation states that for every new element we append to $\mathcal{A}$, we multiply the previous function by the new $\lambda_{D_k,\tilde{S}}$, add a term with the product of the previous λ's multiplied by $(1-\lambda_{D_k,\tilde{S}})$, and multiply the result by (-1). Starting from $G(\emptyset)=0$ and $G(\{S\})=1-\lambda_{D_k,S}$, it follows that the function must take the form

$$G(\mathcal{A}) = (-1)^{|\mathcal{A}|+1} \sum_{S\in\mathcal{A}} (1-\lambda_{D_k,S}) \prod_{S'\in\mathcal{A}\setminus S} \lambda_{D_k,S'} \tag{83}$$

Therefore

$$\begin{aligned}\mathbb{E}(D_K,\mathcal{E})_{\text{dis}} &= \frac{1}{p(\mathcal{S}_+,\mathcal{S}_-|\mathcal{R})} \sum_{\mathcal{A}\subseteq\mathcal{S}_+} \frac{1}{\prod_{S\in\mathcal{A}}\lambda_{D_k,S}} (-1)^{|\mathcal{A}|+1} \sum_{S\in\mathcal{A}} (1-\lambda_{D_k,S}) \prod_{S'\in\mathcal{A}\setminus S} \lambda_{D_k,S'}\; p(\mathcal{S}_-=0,\mathcal{A}=0,D_k=1|\mathcal{R}) \\ &= \frac{1}{p(\mathcal{S}_+,\mathcal{S}_-|\mathcal{R})} \sum_{\mathcal{A}\subseteq\mathcal{S}_+} (-1)^{|\mathcal{A}|+1} p(\mathcal{S}_-=0,\mathcal{A}=0,D_k=1|\mathcal{R}) \sum_{S\in\mathcal{A}} \frac{1-\lambda_{D_k,S}}{\lambda_{D_k,S}} \end{aligned} \tag{84}$$

Once again, we have arrived at a corrected form of the standard posterior

$$\mathbb{E}(D_K,\mathcal{E})_{\text{dis}} = \frac{1}{p(\mathcal{S}_+,\mathcal{S}_-|\mathcal{R})} \sum_{\mathcal{A}\subseteq\mathcal{S}_+} (-1)^{|\mathcal{A}|} p(\mathcal{S}_-=0,\mathcal{A}=0,D_k=1|\mathcal{R})\gamma(\mathcal{A},D_k) \tag{85}$$

where

$$\gamma(\mathcal{A},D_k) = |\mathcal{A}| - \sum_{S\in\mathcal{A}} \frac{1}{\lambda_{D_k,S}} \tag{86}$$

and we recover $\mathbb{E}(D_K,\mathcal{E})_{\text{dis}} = p(D_k=1|\mathcal{E})$ in the limit $\gamma(\mathcal{A},D_k)\to 1$.

Finally, consider that for some $S\in\mathcal{A}$, $\lambda_{D_k,S}=0$. Note that $p(\mathcal{S}_-=0,\mathcal{A}=0,D_k=1|\mathcal{R}) = p(\mathcal{S}_-=0,\mathcal{A}=0|\mathcal{R},D_k=1)p(D_k=1|\mathcal{R})$. If any $\lambda_{D_k,S}=0$ for $S\in\mathcal{S}_-$, then this term is 0 by construction.

□

Proof. Follows immediately from Theorems 4 and 3. Note that this expression is identical to the expression given in Theorem 3 for the expected sufficiency. Hense, in three layer noisy-or models, the expected sufficiency and the expected disablement exactly coincide. □

G. PROPERTIES OF THE EXPECTED DISABLEMENT

In this appendix we show that the expected disablement satisfies our criteria for diagnostic measures. Although in noisy-or networks the expected disablement coincides with the expected sufficiency, which we have already shown to obey our postulates, we show here that the expected disablement in obeys our postulates in general models - regardless of the choice of graph topology or generative functions.

Theorem 6 (Diagnostic properties of expected disablement)**.** *The expected disablement, defined as*

$$\mathbb{E}(D_K,\mathcal{E})_{dis} := \sum_{\mathcal{S}'} \big|\mathcal{S}_+ \setminus \mathcal{S}'_+\big|\, p(\mathcal{S}'|\mathcal{E},do(D_k=0))$$

satisfies the following three conditions

1. consistency. $\mathbb{E}_{dis}(D_K,\mathcal{E}) \propto p(D_k=1|\mathcal{E})$

2. causality. *If* $\nexists\ S\in \mathsf{Dec}(D_k)\cap\mathcal{S}_+ \implies E_{dis}(D_K,\mathcal{E})=0$

3. simplicity. $|E_{dis}(D_K,\mathcal{E})| \le |\mathcal{S}_+\cap\mathsf{Dec}(D_k)|$

Proof. First we prove consistency. In the following, we use the notation $*$ to denote counterfactual variables. The term $p(\mathcal{S}'^{*}|\mathcal{E}, \text{do}(D_k^* = 0))$ can be expressed as

$$p(\mathcal{S}'^{*}|\mathcal{E}, \text{do}(D_k^* = 0)) = \sum_{d_k \in \{0,1\}} p(\mathcal{S}'^{*}, D_k = d_k|\mathcal{E}, \text{do}(D_k^* = 0)) \tag{87}$$

$$= \sum_{d_k \in \{0,1\}} p(\mathcal{S}'^{*}|D_k = d_k, \mathcal{E}, \text{do}(D_k^* = 0))p(D_k = d_k|\mathcal{E}, \text{do}(D_k^* = 0)) \tag{88}$$

As D_k is not a descendent of D_k^*, this simplifies to

$$p(\mathcal{S}'^{*}|\mathcal{E}, \text{do}(D_k^* = 0)) = \sum_{d_k \in \{0,1\}} p(\mathcal{S}'^{*}|D_k = d_k, \mathcal{E}, \text{do}(D_k^* = 0))p(D_k = d_k|\mathcal{E}) \tag{89}$$

If $D_k = 0$ then the factual and counterfactual symptoms have identical states on their parents, and therefor are copies of each other. As a result, $\mathcal{S}_+ = \mathcal{S}'_+$ and the expected disablement is identical to 0. The only term that is non-zero is therefore when $F_k = 1$. Therefore, as $p(D_k = 1|\mathcal{E}) \to 0$, $\mathbb{E}(D_K, \mathcal{E}) \to 0$ smoothly. To see that causality is satisfied, note that $\left|\mathcal{S}_+ \setminus \mathcal{S}'_+\right| \neq 0$ iff $\mathcal{S}'_+ \subset \mathcal{S}_+$, which requires that at least one symptom has been switched off. If D_k is not a parent of any $\mathcal{S}_+$, then $p(\mathcal{S}'^{*}|\mathcal{E}, \text{do}(D_k^* = 0)) = 0$ unless $\mathcal{S}'^{*} = \mathcal{S}$ (the symptom evidence is unchanged), which implies that $\left|\mathcal{S}_+ \setminus \mathcal{S}'_+\right| = 0$, satisfying causality. Finally, note that $\mathbb{E}(D_K, \mathcal{E})$ is a convex combination over the values of the set difference function $\left|\mathcal{S}_+ \setminus \mathcal{S}'_+\right|$, and therefore is upper bounded by $\mathbb{E}(D_K, \mathcal{E}) \leq |\mathcal{S}_+|$, the number of positively evidenced symptoms that are children of D_k. Therefore, the expected disablement is upper bounded by the maximal number of positive symptoms that can be caused by D_k.

□

H. APPENDIX OF EXPERIMENTAL RESULTS

In this appendix we list the results of experiments 1 and 2. Experiment 1 compares the top k accuracy of our algorithms. In experiment 2 the 44 experiments where we pitted doctors against our two diagnostic algorithms. The recorded values are the scores of each agent in each experiment, to three significant figures, with a corresponding standard deviation.

TABLE III: Results for experiment 1: table shows the top k accuracy for the posterior, expected disablement and expected sufficiency ranking algorithms, for N from 1 to 20.

N	Posterior	Disablement	Sufficiency
1	0.496 ± 0.012	0.513 ± 0.012	0.514 ± 0.012
2	0.649 ± 0.011	0.69 ± 0.011	0.686 ± 0.011
3	0.736 ± 0.011	0.776 ± 0.01	0.778 ± 0.01
4	0.782 ± 0.01	0.834 ± 0.009	0.833 ± 0.009
5	0.821 ± 0.009	0.874 ± 0.008	0.874 ± 0.008
6	0.846 ± 0.009	0.901 ± 0.007	0.9 ± 0.007
7	0.867 ± 0.008	0.923 ± 0.006	0.923 ± 0.006
8	0.885 ± 0.008	0.94 ± 0.006	0.938 ± 0.006
9	0.9 ± 0.007	0.952 ± 0.005	0.953 ± 0.005
10	0.912 ± 0.007	0.961 ± 0.005	0.962 ± 0.005
11	0.92 ± 0.006	0.966 ± 0.004	0.967 ± 0.004
12	0.925 ± 0.006	0.972 ± 0.004	0.973 ± 0.004
13	0.933 ± 0.006	0.978 ± 0.003	0.978 ± 0.003
14	0.939 ± 0.006	0.981 ± 0.003	0.981 ± 0.003
15	0.945 ± 0.005	0.986 ± 0.003	0.988 ± 0.003
16	0.95 ± 0.005	0.991 ± 0.002	0.993 ± 0.002
17	0.955 ± 0.005	0.994 ± 0.002	0.995 ± 0.002
18	0.959 ± 0.005	0.997 ± 0.001	0.998 ± 0.001
19	0.963 ± 0.004	0.997 ± 0.001	0.999 ± 0.001
20	0.968 ± 0.004	0.997 ± 0.001	0.999 ± 0.001

TABLE IV: Results for experiment 1: table shows the mean position of the model disease for the associative (A) and counterfactual (C, expected sufficiency) algorithms over all 1763 cases. Results are stratified over the rareness of the disease (given the age and gender of the patient). For each disease rareness category, the number of cases N is given. Also the number of cases where the associative algorithm ranked the model disease higher than the counterfactual algorithm (Wins (A)), the counterfactual algorithm ranked the model disease higher than the associative algorithm (Wins (C)), and the number of cases where the two algorithms ranked the model disease in the same position (Draws) are given, for all cases and for each disease rareness class.

	Cases						
	All	**Very common**	**Common**	**Uncommon**	**Rare**	**Very rare**	**Almost Impossible**
N	1763	139	437	564	375	224	23
Mean position (A)	3.65 ± 4.79	2.90±4.01	2.92± 4.09	3.37± 4.41	4.17 ± 4.92	4.94± 5.68	5.39±6.11
Mean position (C)	2.77±3.12	2.35± 2.65	2.25± 2.68	2.58± 2.73	3.18 ± 3.34	3.71± 4.12	4.13±4.28
Wins (A)	43	6	9	12	14	2	0
Wins (C)	435	23	85	127	109	82	8
Draws	1285	110	343	425	252	140	15

TABLE V: Results for experiment 2: table shows the accuracy obtained by the doctor and each algorithm shadowing the doctors, for each of the 44 single-doctor experiments. The accuracies are reported with the standard standard deviation of the mean estimator.

Doctor number	Doctor accuracy	Posterior accuracy	Expected sufficiency	Expected disablement
0	0.725 ± 0.018	0.658 ± 0.019	0.682 ± 0.019	0.682 ± 0.019
1	0.827 ± 0.022	0.754 ± 0.025	0.781 ± 0.024	0.781 ± 0.024
2	0.894 ± 0.017	0.775 ± 0.024	0.804 ± 0.023	0.804 ± 0.023
3	0.81 ± 0.022	0.778 ± 0.024	0.817 ± 0.022	0.82 ± 0.022
4	0.771 ± 0.034	0.828 ± 0.03	0.892 ± 0.025	0.898 ± 0.024
5	0.616 ± 0.028	0.768 ± 0.024	0.816 ± 0.022	0.819 ± 0.022
6	0.798 ± 0.02	0.748 ± 0.022	0.793 ± 0.02	0.796 ± 0.02
7	0.779 ± 0.026	0.74 ± 0.027	0.782 ± 0.025	0.782 ± 0.025
8	0.692 ± 0.025	0.778 ± 0.023	0.828 ± 0.021	0.834 ± 0.02
9	0.687 ± 0.057	0.746 ± 0.053	0.776 ± 0.051	0.776 ± 0.051
10	0.903 ± 0.038	0.903 ± 0.038	0.935 ± 0.031	0.935 ± 0.031
11	0.777 ± 0.034	0.696 ± 0.038	0.797 ± 0.033	0.797 ± 0.033
12	0.697 ± 0.053	0.724 ± 0.051	0.763 ± 0.049	0.763 ± 0.049
13	0.639 ± 0.061	0.721 ± 0.057	0.754 ± 0.055	0.754 ± 0.055
14	0.802 ± 0.031	0.76 ± 0.033	0.76 ± 0.033	0.76 ± 0.033
15	0.888 ± 0.018	0.745 ± 0.024	0.776 ± 0.023	0.776 ± 0.023
16	0.8 ± 0.041	0.811 ± 0.04	0.863 ± 0.035	0.874 ± 0.034
17	0.652 ± 0.051	0.753 ± 0.046	0.753 ± 0.046	0.753 ± 0.046
18	0.736 ± 0.042	0.809 ± 0.037	0.827 ± 0.036	0.827 ± 0.036
19	0.796 ± 0.025	0.737 ± 0.028	0.796 ± 0.025	0.8 ± 0.025
20	0.712 ± 0.03	0.673 ± 0.031	0.717 ± 0.03	0.718 ± 0.03
21	0.697 ± 0.028	0.675 ± 0.028	0.745 ± 0.026	0.753 ± 0.026
22	0.718 ± 0.034	0.655 ± 0.036	0.724 ± 0.034	0.714 ± 0.034
23	0.614 ± 0.046	0.675 ± 0.044	0.711 ± 0.042	0.704 ± 0.043
24	0.698 ± 0.063	0.679 ± 0.064	0.717 ± 0.062	0.704 ± 0.062
25	0.852 ± 0.033	0.783 ± 0.038	0.817 ± 0.036	0.826 ± 0.035
26	0.807 ± 0.036	0.773 ± 0.038	0.849 ± 0.033	0.832 ± 0.034
27	0.684 ± 0.044	0.658 ± 0.044	0.746 ± 0.041	0.728 ± 0.042
28	0.579 ± 0.044	0.651 ± 0.042	0.659 ± 0.042	0.664 ± 0.042
29	0.757 ± 0.05	0.811 ± 0.046	0.892 ± 0.036	0.892 ± 0.036
30	0.813 ± 0.038	0.682 ± 0.045	0.738 ± 0.042	0.729 ± 0.043
31	0.523 ± 0.062	0.585 ± 0.061	0.615 ± 0.06	0.606 ± 0.06
32	0.631 ± 0.06	0.692 ± 0.057	0.723 ± 0.056	0.712 ± 0.056
33	0.606 ± 0.06	0.606 ± 0.06	0.667 ± 0.058	0.657 ± 0.058
34	0.585 ± 0.061	0.631 ± 0.06	0.662 ± 0.059	0.652 ± 0.059
35	0.641 ± 0.06	0.641 ± 0.06	0.703 ± 0.057	0.692 ± 0.057
36	0.601 ± 0.037	0.567 ± 0.037	0.624 ± 0.036	0.624 ± 0.036
37	0.706 ± 0.044	0.734 ± 0.042	0.798 ± 0.038	0.789 ± 0.039
38	0.694 ± 0.044	0.741 ± 0.042	0.824 ± 0.037	0.824 ± 0.037
39	0.769 ± 0.058	0.769 ± 0.058	0.865 ± 0.047	0.865 ± 0.047
40	0.628 ± 0.05	0.787 ± 0.042	0.84 ± 0.038	0.851 ± 0.037
41	0.747 ± 0.048	0.614 ± 0.053	0.711 ± 0.05	0.72 ± 0.05
42	0.55 ± 0.064	0.633 ± 0.062	0.7 ± 0.059	0.7 ± 0.059
43	0.746 ± 0.053	0.791 ± 0.05	0.791 ± 0.05	0.791 ± 0.05