Reprinted from Proceedings of the Fifteenth Annual Symposium on Computer Applications in Medical Care, pp 619-623, 1991

The Probability of Disease*

William J. Long
MIT Laboratory for Computer Science
Cambridge, MA 02139

This paper addresses the nature of the prior probabilities of diseases for probabilistic diagnostic reasoning. Because diseases differ in their chronicity, occurrence, reoccurrence, and likelihood of becoming part of the patient population, reasoning in terms of the frequency of disease episodes is necessary to capture the important distinctions. Even with these complexities, it is possible to formulate a reasonably accurate, computationally tractable, frequency estimation method for combinations of diseases. This method also suggests ways in which the needed numbers can be estimated from patient data.

1 Introduction

Many medical diagnosis programs use probabilistic inference to reason about the attribution of findings (for example [1, 2, 3]). The usual paradigm is that diseases have some prior probability of occurring and produce findings with some probability, possibly through some intermediate states. Thus, the view of the diagnosis problem corresponds well to reasoning about Bayesian probability networks. A fundamental assumption is that the primary causes, the diseases, are independent. If the diseases are known not to be independent, etiologies can be added to the network as primary causes to eliminate the dependencies. This representation of the diagnostic problem has great intuitive appeal, but a number of difficulties arise when applying it to real medical domains. The problem addressed by this paper is why one often finds patients with several seemingly independent diseases when the product of the probabilities would predict that this would be a rare event.

Consider the situation where disease A has a prior probability of 0.01 and disease B has a prior probability of 0.05. Does that mean that the next patient in the door has probabilities of 0.01 of having A, 0.05 of having B, and 0.0005 of having both? The answer is no for several reasons. 1) People rarely come into a hospital admitting door (or doctors office, or appropriate context for the medical domain) unless they think they have something that needs attention. Thus, the probability of some disease in the patient is high and certainly much higher than in the man on the street. 2) A or B may not always cause the patient to go to the hospital. 3) The probability of the combination of the diseases, assuming that they are causally independent, still depends on how long each one persists. If both A and B have acute onset and do not remain long (like appendicitis, for example), it is very unlikely that they will occur together. If one or both is a chronic disease, it is much more likely that the patient who comes into the hospital will have both.

2 The Nature of the Problem

The basic problem is the meaning of the probability of disease. That is, the meaning of the numbers, often called *priors*, that are assigned to the primary disease entities. To have a probability there must be a defined population over which it is determined. A simple answer might be to assume the population is that served by the hospital (or care unit). This model serves well for a program such as deDombal's acute abdomen program[4], where the diseases in question are acute events, typically occur once in a lifetime, and always cause the patient to go to the hospital. Variations from these characteristics cause problems. If one considers the probability of influenza, it becomes clear that the issue is not the number of patients with the disease, but the frequency of the disease. Since each person typically has influenza several times during his or her lifetime, it does not make sense to talk about the probability of the disease. It makes more sense to say that the average frequency of influenza is once in ten years. Other diseases are chronic and present another set of issues. If a patient becomes diabetic, they remain diabetic for the rest of their lifetime. It makes sense to say that there is a certain probability that a person will be diabetic by the end of their life, but that obscures the problem since the patient can

*This research was supported by National Institutes of Health Grant No. R01 HL33041 from the National Heart, Lung, and Blood Institute and No. R01 LM04493 from the National Library of Medicine.

enter the hospital many times as a diabetic or before they become diabetic. The context in which the occurrence of disease is of concern is the hospital visit. Therefore, a typical person with diabetes will account for several "patients", that is, several hospital visits. Thus, the number of concern is the expected frequency of hospital visits for a particular disease.

The example of influenza also raises the issue of whether the person with a disease will become a patient. We could say that we are only interested in diseases severe enough to require a hospital visit, but that is not sufficient. If the patient is otherwise healthy, influenza is unlikely to require a hospital visit, but if the patient has a chronic disease that saps their strength, it is much more likely to require a hospital visit. Thus, even though the two diseases may have independent etiologies, the probability that a person becomes part of the patient population may be dependent on other diseases. Accounting for this phenomenon is particularly important if the domain is limited to a particular specialty. For example, the frequency of patients with pneumonia alone in a cardiology clinic is small, because those patients are treated elsewhere. However, the frequency of pneumonia on top of an existing cardiac problem is high, because pneumonia tends to decompensate the cardiac problems.

3 Example

The basic proposal is to view the problem in terms of the frequency of disease events in the patient. To see the practical implications, consider a population of 1000 people and two possible diseases A and B. Disease A is a chronic incurable disease that occurs in one out of a hundred people. Once a person has it, they average a hospital visit for a "flare-up" every five years and they have an average remaining lifespan of twenty years. Disease B is an acute disease that always sends a person to the hospital but with proper hospital treatment is quickly resolved. It occurs in the average person once every forty years. Consider the expected hospitalizations over the lifetimes of all of the people in the population, assuming that the average lifetime in the population is 80 years. For disease A, the 10 people that contract it will each average four hospitalizations for a total of 40 hospitalizations caused by disease A. For disease B, the average person will have it twice for a total of 2000 hospitalizations. There are also 200 person years when people have disease A and could also contract disease B. Thus, there are 5 expected hospitalizations caused by disease B in which disease A is present. If "flare-ups" of A and incidents of B are considered to have no time duration, the possibility of both happening together can be ignored. Thus, there are 45 hospitalizations of patients with disease A, 11% of which are caused by disease B.

Consider that for disease B the chances of being hospitalized for the disease by itself is 0.1, but if it occurs when disease A is present, the patient is always hospitalized. Now there are 200 hospitalizations for B, but there are still 5 for the combination of A and B.

4 Frequency of Patients with a Disease in a Setting

Given that the number to calculate is the expected frequency of patients with the disease in the particular care setting, there are several factors that go into that calculation. 1) The probability that a patient will contract the disease during a given time period. This number may be dependent on age, sex or other demographic factors, but should be independent of any other disease except where the dependency is explicitly represented. 2) The time course of the disease. In the domain of cardiovascular disorders that we have been studying, there are two primary types of time course, diseases that are acute, requiring a single admission, and ones that are permanent unless surgically corrected. There are some other time courses in other domains, but these cover the main issues. Acute diseases can usually be considered to be events at a point in time. As such, the probability of two such diseases happening at the same time in the patient is vanishingly small unless there is a causal relation between them. The practical implications of this assumption needs to be carefully considered for each domain, otherwise important situations may be ruled out. For example, the causal relation between diseases may be vague — influenza can cause a myocardial infarction by putting stress on a heart with preexisting coronary artery disease. Also, the hospital stay for an acute disease may be long enough for the patient to contract a second acute disease, such as pneumonia. This is particularly important when it is the second disease that brings the patient into the hospital. Chronic diseases can be considered always present. 3) For chronic diseases, it is necessary to know the expected frequency of significant episodes and the effect of the disease on the patient's life expectancy. That is, if a person contracts a chronic disease at age 40 with yearly episodes requiring hospitalization but with no change in life expectancy, there will be many admissions with that disease. If however, the average life expectancy after contracting the disease is a year, there will be few admissions. The frequency of episodes of a disease may be dependent on many factors, but it is difficult to get sufficient data on particular diseases to characterize these relations in much detail. 4) The final factor is the likelihood that the patient with a disease episode will become part of the cohort of interest. This is usu-

ally dependent on the overall state of the patient.

If these factors can be specified for the diseases of concern, the next step is to compute the expected frequency of disease. Consider first the single disease situation for a patient of age y. For an acute disease contracted on average $q(y)$ times per patient-year and requiring admission h fraction of the time, the expected frequency is $q(y) \times h$.

For a chronic disease, the calculation is more complicated. The patient contracted the disease at some time, survived another a years and had an episode at age y. Since the frequency is relative to all patients who attained age y, the survival needs to be relative to normal survival. Since all such patients are included, the function is summed over all ages less than y. An approximation of the change in survival that is reasonably accurate for a wide range of diseases and is computationally tractable is the declining exponential[5]. That is, the size of the population of patients who contracted the chronic disease changes relative to the population of patients without the disease over the a years at e^{-ra}. Another way to think about this equation is in terms of the number of years it would take to kill half of the patients that would have remained alive without the disease. If a is that number of years, $r = .693/a$.

In the simplest case where the incidence of disease q is constant for age, the probability of hospitalization h is constant, and the frequency of episodes f is independent of both age and length of time the disease has been present, the expected frequency of episodes is obtained by summing over a from 0 to age y:

$$\int_0^y qfhe^{-ra}da = qfh\frac{1-e^{-ry}}{r}$$

The fraction represents the average number of years the patient would have had the disease. For example, if the patient is 50 years old and r is 0.05, reflecting a five year diminished survival of 78%, the expected length of time the patient had the disease is 18 years. If there is an episode on average every 5 years, this will be (on average) the third or fourth episode. If the yearly risk of contracting the disease is 0.01 and an episode always causes a hospital admission, the expected frequency of hospital admissions is $0.01 \times 0.2 \times 1.0 \times 18.4 = 0.037$ per person per year.

For simple dependencies of q on age, it is possible to generalize this relation. For example, if there is no risk of the disease until a certain age b and then the yearly risk is q, the frequency becomes

$$qfh\frac{1-e^{-r(y-b)}}{r}.$$

If there is a change in the risk at some age from q to kq, the frequency is

$$\frac{qfh}{r}(1-\frac{k-1}{k}e^{-r(y-b)}-\frac{1}{k}e^{-ry}).$$

Often, the frequency of episodes of a chronic disease increase over time. One way to model this behavior is with the frequency as fe^{ka}. That is, initially the frequency is f but that doubles in $\log 2/k$ years. The expected frequency of episodes would be

$$\int_0^y qfe^{ka}he^{-ra}da = qfh\frac{1-e^{(k-r)y}}{k-r}.$$

Care must be taken in representing a disease this way because the frequency of episodes never gets above about half a dozen times a year before the episodes merge into a single hospital stay or the patient succumbs to the disease. It is possible for the k to be larger than the r exponent. For example, for a disease that has an excess mortality of 20% each five years, the r is about 0.05. If the frequency of hospital episodes increases from one every four years to four a year over thirty years with the disease, the k is 0.09. This simply means that the 22% of the patients (adjusting for normal life expectancy) remaining after 30 years with the disease are responsible for more episodes than the initial group when they contracted it.

Thus, this scheme is able to approximate the kind of characteristics that might be known about a disease and accounts for the fundamental differences between acute and chronic diseases.

5 Frequency of Two Diseases

The problem we started with is how to estimate the expected frequency of multiple diseases in the patient and we now have the machinery to address that problem. The two cases of concern are an acute disease and a chronic one, and two chronic diseases. As discussed above, the probability that two acute diseases happen together is zero unless there is a causal relationship between them. If this is a problem, one could assign time extents to them. For example, just multiplying the frequencies together is the frequency that they happen in the same year. An alternative solution would be to include a nonspecific causal mechanism that can cause certain acute diseases when the patient is already "sick" with an acute disease. In each domain the knowledge engineer needs to ascertain the significance and appropriate model for acute diseases that happen together.

5.1 An Acute and a Chronic Disease

When the patient has both an acute and a chronic disease, the acute disease struck during the tenure of

the chronic disease. There is not a coincidence of an episode of the chronic disease with the acute disease, for the same reasons that acute diseases do not happen together. Thus, the patient has contracted the chronic disease c, had it for some period of time, contracted the acute disease a, and has been hospitalized by the combination. The frequency with which that occurs is

$$q_c \frac{1 - e^{-r_c y}}{r_c} q_a h_{a|c}.$$

The only part of this equation that is unknown is the probability $h_{a|c}$ of being hospitalized with the acute disease given the chronic one. Under almost all situations it will be higher than the probability of being hospitalized with the acute disease alone, but it may have nothing to do with the probability of being hospitalized with an episode of the chronic disease, since there is no episode. A plausible mechanism for estimating the probability would be to have an additional factor for chronic diseases indicating the degree to which the disease in general "compromises" the patient to be combined with the probability of hospitalization for the acute disease. However, one can think of situations where this would not be satisfactory and more empirical experience is needed to develop an appropriate mechanism for handling this in general.

5.2 Two Chronic Diseases

When the patient has two chronic diseases, one is undergoing an episode, but not both. Fortunately, part of the diagnosis is which disease is having an episode. The situation is that the patient contracted chronic disease 1, some time passed, contracted chronic disease 2, some time passed, is now experiencing an episode of 1, and is hospitalized. Alternatively, disease 2 preceded disease 1. The first expected frequency of the first situation is

$$q_1 e^{-r_1(b-a)} q_2 e^{-(r_1+r_2)(y-b)} f_1 h_{1|2}.$$

The frequency of this happening for all times a and b is

$$\int_0^y \int_0^b q_1 e^{-r_1(b-a)} q_2 e^{-(r_1+r_2)(y-b)} f_1 h_{1|2} da\, db.$$

Integrating and adding the corresponding term for disease 2 preceding disease 1 gives a formula for the expected frequency of episodes of disease 1:

$$q_1 \frac{1 - e^{-r_1 y}}{r_1} q_2 \frac{1 - e^{-r_2 y}}{r_2} f_1 h_{1|2}.$$

This expression can be grouped into four parts, the expected years of diseases 1 and 2, the frequency of episodes and the probability of hospitalization. The probability of hospitalization depends on the episode of disease 1 in the context of both diseases. If there is some way to characterize the degree of compromise of the two diseases, this might be the probability implied by an episode of disease 1 combined with the compromise of disease 2.

If the frequencies of episodes of the diseases are modeled as increasing, the expected frequency is

$$q_1 \frac{1 - e^{(k_1 - r_1)y}}{k_1 - r_1} q_2 \frac{1 - e^{-r_2 y}}{r_2} f_1 h_{1|2}.$$

6 Application to Medical Domains

Given that this representation of patient hospitalizations in terms of frequency of disease episodes makes it possible to provide appropriate estimates of the expected frequency of different disease combinations, the remaining question is how to get the needed numbers. There are several kinds of information available in hospital records about the occurrence of diseases. For each admission there is information about the primary problem and a list of the other problems that the patient had. In addition, one can look back in the patient record or statements of the patient's history and get a pretty good indication of how many times and when the patient was admitted in the past for the same problems. This information is less reliable because there can be omissions for a number of reasons. If one has the number of past admissions for a particular disease and the number of years the patient has had that disease, the frequency of disease episodes can be estimated directly (with appropriate corrections for estimating the beginning of the disease from the first episode).

The rest of the numbers can be estimated from disease combinations. If we have data covering a period of time, the number of episodes c of a particular acute disease is an estimate of nph where n is the size of the patient population. If the statistics cover all of the instances of the disease, h can be assumed to be one. The population size is the same for all of the diseases. For chronic diseases, let

$$p = q(\frac{1 - e^{-ry}}{r}).$$

Then, the count of episodes of the disease is an estimate of nfp (assuming $h = 1$). Since p is a function of the age of the patient, age ranges need to be considered. The p's can be estimated by looking at disease combinations. The expected number of acute disease 1 in patients with chronic disease 2 is np_1p_2. Thus, the ratio of the counts of the combination to the count of the acute disease $c_{1|2}/c_1$ estimates p_2 (of the

chronic disease). Similarly, the count of admissions for episodes of chronic disease 1 when chronic disease 2 is also present is an estimation of $nf_1p_1p_2$, so the ratio of counts $c_{1|2}/c_1$ estimates p_2. In general, all episodes of other diseases estimate the p_i of a chronic disease, therefore

$$p_i \approx \frac{\sum_j c_{j|i}}{\sum_j c_j},$$

where all of the diseases j are ones in which the fraction admitted is independent of disease i (ie, $h_{j|i} = h_j$).

Once the p's of chronic diseases have been estimated, the population size can be estimated and the probabilities of acute diseases chosen accordingly. The p's of chronic diseases also give another way of estimating the frequency of episodes if that frequency is independent of the length of time the patient has had the disease. If not, a more thorough analysis of the patterns of episode frequency must be done.

7 Summary

This paper presents a solution to the problem of estimating the likelihood of combinations of diseases in the patient. The problem is recast in terms of the expected frequency of the disease episodes that would be encountered in a particular setting. This clarifies the nature of the calculation and makes it possible to account for the important characteristics of the problem. The solution is practical because the factors needed to do the estimated frequency calculations can be estimated from the kind of data normally available about diseases.

References

[1] William Long, "Medical Diagnosis Using a Probabilistic Causal Network," *Applied Artificial Intelligence,* **3:** 367-383, 1989.

[2] K. G. Olesen, U. Kjaerulff, et al, "A Munin Network for the Median Network — A Case Study on Loops," *Applied Artificial Intelligence,* **3:** 385-403, 1989.

[3] M. Shwe, B. Middleton, et al, "A Probabilistic Reformulation of the Quick Medical Reference System," *14th SCAMC,* Washington, DC, pp790-794, 1990.

[4] F. T. deDombal, D. J. Leaper, et al, "Computer-Aided Diagnosis of Acute Abdominal Pain," *British Medical Journal,* **2:** 9-13, 1972.

[5] J. Robert Beck, Jerome P. Kassirer, and Stephen G. Pauker, "A Convenient Approximation of Life Expectancy (The 'DEALE')," *The American Journal of Medicine,* **73:** 883-897, 1982.