# MODELLING THE PATHOGENESIS OF CYSTIC FIBROSIS AND OTHER MONOGENIC CONDITIONS, AND THE OCCURRENCE OF CAUSATIVE VARIANTS

by

**Johan Willie Viljoen**

Submitted in partial fulfilment of the requirements for the degree
Doctor of Philosophy (Electronic Engineering)

in the

Department of Electrical, Electronic and Computer Engineering
Faculty of Engineering, Built Environment and Information Technology
and
Department of Immunology
Faculty of Health Sciences

UNIVERSITY OF PRETORIA

August 2019

…altyd Antjé

# SUMMARY

## MODELLING THE PATHOGENESIS OF CYSTIC FIBROSIS AND OTHER MONOGENIC VARIATIONS, AND THE OCCURRENCE OF CAUSATIVE VARIANTS

by

**Johan Willie Viljoen**

| | |
|---|---|
| Supervisors: | Prof. J.P. de Villiers and Prof. M.S. Pepper |
| Departments: | Electrical, Electronic and Computer Engineering |
| | Immunology |
| University: | University of Pretoria |
| Degree: | Philosophiæ Doctor (Electronic Engineering) |

Keywords: Mutations, population genetics, monogenic disorders, cystic fibrosis, numeric simulations, graphical models, Bayesian networks, diagnostic support, recessive alleles, balancing selection

Deleterious recessive monogenic autosomal conditions are modelled both on an individual level, for diagnostic purposes, as well as in large populations, where the establishment, dispersion and equilibrium behaviour is investigated.

Data fusion techniques are applied to combine diagnostic data on a more rigorous basis, to support the diagnosis of disease in an individual. In this case the focus is specifically on cystic fibrosis, which is one of the most common monogenic recessive disorders in humans.

Diagnostic information may be of disparate types and varying verisimilitude, such as symptoms, measurements, history, observations, and even opinions. Nonetheless it is possible to construct a mathematical framework to synthesise this knowledge into a numeric assessment of the probability that the disease may be present. This may be used to guide decisions regarding treatment or additional testing, by supporting improved cost-benefit analyses.

Considering the population genetics of monogenic variations such as cystic fibrosis, analytical and statistical stochastic approaches are used to model and predict the dispersion of mutations through a large population. These approaches are used to quantify the magnitude of a heterozygous selective advantage of a mutation in the presence of a homozygous disadvantage. Random effects such as genetic drift are accounted for, which are likely to extinguish even highly advantageous mutations while the prevalence is still low. Dunbar’s results regarding the cognitive upper limit of the number of stable social relationships that humans can maintain are used to determine a realistic community size - a reduced local subset of the total population - from which an individual can select mates. This reduction has a dramatic effect on the probability of establishing mutations, as well as the eventual equilibrium values that are reached in the case of mutations conferring a heterozygous selective advantage, but a homozygous disadvantage, as in the case of cystic fibrosis and sickle cell disease. The magnitude of this selective advantage can then be estimated based on observed occurrence levels of a specific mutation in a population, without requiring prior information regarding its phenotypic manifestation.

It is also demonstrated that the heterozygous carrier levels of monogenic recessive disorders are routinely overestimated.

# OPSOMMING

## MODELLERING VAN DIE ONTLUIKING VAN SISTIESE FIBROSE EN ANDER MONOGENIESE AFWYKINGS, ASOOK DIE VOORKOMS VAN OORSAAKLIKE VARIANTE

deur

**Johan Willie Viljoen**

| | |
|---|---|
| Studieleier: | Prof J.P. de Villiers en Prof M.S. Pepper |
| Departemente: | Elektriese -, Elektroniese - en Rekenaaringenieurswese |
| | Immunologie |
| Universiteit: | Universiteit van Pretoria |
| Graad: | Philosophiæ Doctor (Elektroniese Ingenieurswese) |
| Sleutelwoorde: | Mutasies, bevolkingsgenetika, monogeniese afwykings, sistiese fibrose, numeriese simulasies, grafiese modelle, Bayesnetwerke, diagnostieke steun, resessiewe gene, ewewig |

Nadelige resessiewe monogeniese afwykings word bestudeer in individue, vir diagnostieke doeleindes, asook in groot bevolkings, waar vestiging, verspreiding en ewewigsgedrag ondersoek word.

Datafusietegnieke word ingespan om diagnostieke inligting op 'n meer objektiewe wyse te kombineer ter ondersteuning van die diagnose van siekte in 'n individu. In hierdie geval word spesifiek gefokus op sistiese fibrose, wat die mees algemene menslike monogeniese resessiewe afwyking is. Die insetdata kan van uiteenlopende aard wees, met wisselende

grade van geloofbaarheid, soos simptome, metings, waarnemings, geskiedenis, en selfs vermoedens. Desnieteenstaande is dit moontlik om 'n wiskundige raamwerk te bou wat al hierdie inligting kan kombineer tot 'n numeriese raming van die waarskynlikheid dat die toestand teenwoordig mag wees. Dit kan dan gebruik word om besluite rakende behandeling of bykomende toetse te rig, deur verbeterde koste-voordeel-ontledings moontlik te maak.

Hierna word die bevolkingsgenetika van monogeniese variasies gemodelleer. Oënskynlik nadelige mutasies soos sistiese fibrose en sekelsel-anemie versprei deur menslike bevolkings danksy die selektiewe voordeel wat dit aan heterosigotiese draers bied, onderhewig aan omgewingstoestande. Namate die voorkoms van so 'n mutasie toeneem, verhoog die waarskynlikheid op homosigotiese nageslag daarmee saam, met die geassosieerde selektiewe nadeel wat daarmee gepaard gaan, totdat daar 'n ewewig bereik word tussen draers en nie-draers. Hierdie navorsing gebruik analitiese en stogastiese tegnieke om die absolute grootte van die heterosigotiese voordeel wat so 'n mutasie bied te kwantifiseer. Die antropologiese waarnemings van Dunbar rakende die perke op die menslike vermoë tot stabiele sosiale verhoudings, gekombineer met sensusdata oor die werklike voorkoms van draers, is al wat nodig is om 'n afskatting van die selektiewe heterosigotiese voordeel te maak, sonder om enige voorkennis rakende die fisiese meganismes waardeur so 'n voordeel gebied word te vereis. Dieselfde model kan dan ook gebruik word om die effek van variasies in effektiewe bevolkingsgrootte en selektiewe voordeel op die waarskynlikheid dat 'n variasie hoegenaamd gevestig sal raak te kwantifiseer.

Dit word ook aangetoon dat die draervlakke van monogeniese resessiewe afwykings as 'n reël oorskat word.

## LIST OF ABBREVIATIONS

| | |
|---|---|
| AI | Artificial intelligence |
| BN | Bayesian network |
| CF | Cystic fibrosis |
| CFTR | CF Transmembrane Conductance Regulator |
| ECFS | European Cystic Fibrosis Society |
| IEEE | Institute of Electrical and Electronic Engineers |
| LCT | lactase |
| MHC | Major histocompatibility complex |
| QMR-DT | Quick Medical Reference – Decision Theoretic |
| Rh | Rhesus |
| RTI | Respiratory tract infection |
| UK | United Kingdom |
| USA | United States of America |

# TABLE OF CONTENTS

# LIST OF FIGURES

# CHAPTER 1 INTRODUCTION

## 1.1 PROBLEM STATEMENT

### 1.1.1 Context of the problem

Disease diagnosis is a process whereby a number of observations are combined to determine a most likely finding, which is then used to inform treatment decisions. All observations display some uncertainty or error rate, which may be difficult or impossible to quantify, and hence the eventual verdict can itself never be completely certain. Nonetheless, various types of evidence are combined until a sufficiently high confidence level is reached. The required confidence level itself depends on the impact (i.e. cost and/or risk) of the testing or treatment decisions that it may prompt. Medical practitioners do this routinely, rapidly, largely subconsciously, and possibly not optimally in all cases.

Cystic fibrosis (CF), one of the most common monogenic recessive disorders in humans [1], will be used as a test case to explore evidence combination approaches, to probe the impact of uncertainty on the outcome, and also to characterise the contribution of various observations to the final conclusion. Additionally, CF is an attractive candidate for population modelling purposes, due to its highly deleterious character, which until very recently all but guaranteed that homozygous individuals would produce no progeny.

Observing the relative prevalence of ostensibly deleterious genetic variations such as CF and sickle cell anaemia leads to the inevitable conclusion that such variations *have* to offer some selective advantage, lest they rapidly be outcompeted by the wild type. Sometimes such advantages are known [2]: for example, mutations in the haemoglobin gene can confer enhanced malaria resistance to heterozygous carriers [3], [4], but the outcome is dire for homozygous individuals [5]. This limits the prevalence of the allele. However, many variations are not quite as dramatic in either their heterozygous advantages or homozygous disadvantages, and may be highly dependent on unknown epistasis and epigenetic factors. Nonetheless, observations regarding occurrence can be made. This, combined with sociological factors, will be used to explore the characteristics of recessive deleterious genetic variations in human populations, specifically the likelihood of establishment, the eventual equilibrium prevalence levels, updated carrier level estimates, and even the likely heterozygous selective advantage that it confers, without requiring any information regarding its phenotypic manifestation.

### 1.1.2 Research gap

Current guidelines for the diagnosis of cystic fibrosis suggest a sweat chloride test, followed by genetic screening [6]. However, as shall be shown, the rarity of the condition may lead to high false positive rates (relative to the incidence of CF), which in turn "can have psychosocial consequences that affect entire families" [7]. Furthermore, genetic screening is not always available or completely relevant, especially in Africa [8]. CF presents with a wide range of variable and often non-specific signs and symptoms [1], which suggests that a more objective combination of evidence, based on Bayesian theory, may improve accuracy and specificity.

Considering the establishment and dispersion of monogenic variations in a population, much effort has been expended to analytically model the effects of local structure, migration, and inbreeding [9], [10], [11], [12]. However, there is a strong social dimension to human populations, as in most primates [13], [14] – this affects group structure, and hence also procreation patterns. Until now this effect has not explicitly been taken into account in population genetics models.

## 1.2 RESEARCH QUESTIONS

Many diseases present with non-specific symptoms, which are merged by a medical practitioner to arrive at a diagnosis, which is inevitably at least partially subjective. Because rare diseases are particularly challenging to diagnose, the following questions are considered:

- Can a Bayesian network (BN) be used to gain some insights into the ætiology and pathology of CF?
- How does uncertainty in the inputs affect the utility of a BN used for predicting CF?

At the population level, monogenic variations in humans are studied. Prompted by the research gap identified above, the following questions were considered:

- Why do ostensibly deleterious genetic variations not become extinct?
- Can human sociological data be used to derive improved values of the heterozygous advantage and carrier prevalence associated with homozygous deleterious mutations?
- Can sociological data be used to determine the probability that a genetic variation will become established in the human population?

## 1.3 APPROACH

With CF being one of the most common monogenic recessive disorders in humans, this disease is used as a primary focus area. A BN is created, to take into account various associated risk factors and known symptoms to calculate a probability that a given individual is indeed afflicted with homozygous CF.

Then, using known incidence data and realistic community size numbers, a numeric population model is created to estimate the heterozygous carrier prevalence as well as the selective advantage that a mutation in the CF Transmembrane Regulator (CFTR) gene bestows on heterozygotes.

## 1.4 RESEARCH GOALS

The aim is to create tools to:

- improve insight into the diagnostic process for CF, that may be used to guide decisions regarding tests most likely to add useful data.
- infer improved estimates of characteristics of CF and other monogenic disorders by exploiting sociological and census data.

## 1.5 RESEARCH CONTRIBUTION

The creation of an improved general approach to disease modelling as proposed above may result in a tool which can be of use to health administrators and policy makers, by facilitating causal predictions regarding the expected effects of interventions, as well as making possible improved risk/benefit analyses, which is always relevant when limited resources need to be applied to achieve the greatest good.

It is believed that the incorporation of heterogeneous features exhibiting different types of uncertainty, as well as a more detailed longitudinal modelling than hitherto reported in the literature may lead to a significantly improved result.

On the level of individuals rather than groups, a sufficiently detailed and accurate disease model would make truly personalised medicine feasible, by allowing the specific factors (genetic or environmental) besetting the patient to be analysed in their complex interactions; selection and adjustment of an intervention to the best known level possible for that patient then becomes viable.

Furthermore, and more directly, it is expected that the envisaged cystic fibrosis case study will result in useful insights regarding improved management of this disease, which should lead to better resource management, as well as improved quality of life and/or increased lifespan for those afflicted.

The combination of sociological and census data in a stochastic model as applied to the population genetics of monogenic variations adds a hitherto overlooked dimension to the existing body of theory, and is shown to reproduce various known boundary test cases with remarkable accuracy, while at the same time obviating the need for a number of postulated and usually unknown confounding factors traditionally used in genetic models of populations. This model, when set to replicate known CF incidence levels, generates estimates of the heterozygous selective advantage and also indicates a significant downward adjustment in current carrier prevalence estimates for autosomal monogenic recessive disorders.

## 1.6 RESEARCH OUTPUTS

Journal paper published in *Nature Scientific Methods* on 2019-07-17 [15].

## 1.7 OVERVIEW OF STUDY

In Chapter 2 the results of two literature reviews are presented, respectively investigating data fusion for medical decision support (section 2.2), and the population genetics of monogenic variations (section 2.3). In both cases, while a generic viewpoint is presented, the applicability to deleterious recessive monogenic variations, and specifically cystic fibrosis, is consistently kept in mind. It is found that medical decision support is a fertile area especially for artificial intelligence (AI) research, but that despite decades of promising results, there is still a paucity of successful applications, due to a number of practical reasons.

Regarding the dissemination of monogenic mutations in large populations, an impressive body of experimental data exists, especially on fruit flies, but for obvious reasons human experimentation is fraught with practical and ethical limitations, spurring the development of analytical models to predict the behaviour of variations in large homogeneous and structured populations.

Chapter 3 then presents the research methods that were applied to disease modelling, specifically aimed at the modelling of cystic fibrosis as a test case, by using Bayesian networks to integrate environmental, physiological, genetic, historical and even unknown (postulated) evidence to arrive at an improved estimate of the likelihood that disease is indeed present in an individual.

While Chapter 3 focuses on disease in an individual, Chapter 4 investigates cystic fibrosis and similar monogenic variations from a group (population) perspective, and demonstrates that incorporation of human sociological data into an essentially Mendelian stochastic model can reproduce the results of rather abstruse analytical approaches, while obviating the need to estimate several unknown factors, the effects of which appear as emergent features instead. It is shown that plausible estimates can be obtained for a fundamental characteristic of cystic fibrosis, namely the heterozygous selective advantage, as well as the heterozygous carrier frequency, by replicating the known (i.e. observed) homozygous incidence levels.

Chapter 5 presents a conclusion, with several suggestions for future study.

# CHAPTER 2 LITERATURE REVIEW

## 2.1 CHAPTER OBJECTIVES

This chapter is divided into two major parts – the first presents the results of a literature study aimed at determining the current state of medical decision support, while the second focuses on the theory of population genetics, specifically the establishment and dissemination of variations (mutations) in diploid populations.

## 2.2 MEDICAL DECISION SUPPORT FOR RISK DETERMINATION, DIAGNOSIS AND TREATMENT

### 2.2.1 Introduction

In 1885 Sir Arthur Conan Doyle observed in a short story that "*knowledge begets knowledge, as money bears interest*" [16]. This sentiment, if true, would imply exponential growth in knowledge (or at least information), and indeed there have been several studies supporting this theory. David T. Durack in 1978 published the results of a project in which the growth of the *Index Medicus* – a database of biomedical journal articles, maintained by the United States National Library of Medicine – was determined by simply weighing the tomes that had been published over the course of the preceding century [17]. He extrapolated the weight (which had stood at about 3kg in 1927 and ten times as much 50 years later) to reach a predicted 1000kg by 1985, and concluded that the printed format would inevitably have to make way for something like microfilm or "*computer-terminal display*".[1] Others have found similar trends – Price [18] also looked at indicators like the number of scientists, abstracts and journals as well as scientific research expenditure, and concluded that over the preceding two centuries there had been an annual growth rate of approximately 5% in the scientific literature, far outstripping even human population growth figures.

It is of course an oversimplification to equate the amount of published information directly with a growth in knowledge, partly because much of published science also represents correction or even replacement of previous wisdom. Ramsey *et al* in 1991 estimated a half-life of five years for internal medicine knowledge [19]. Less quantitatively, a quotation, popular at higher education institutions (especially medical schools), states that "*half of what we know (or will teach you) is incorrect, but we unfortunately do not know which half.*"[2]

The foregoing clearly demonstrates that scientific knowledge, and specifically medical knowledge, is in a continuous state of flux. While advances are potentially good news for the patient, the current knowledge growth and change rates are such that no physician can realistically be expected to be properly up-to-date on anything except possibly the narrowest of specialisations. This fact all but guarantees that medical diagnoses will be based on less than the current state-of-the-art information. Newman-Toker and Pronovost in 2009 identified diagnostic errors as "*the next frontier for patient safety*" [21], underlining the need for diagnostic decision support tools that can access and meaningfully utilize more of the latest knowledge than the physician can.

[1] Although growth in the weight of the printed *Index Medicus* had slowed down somewhat in the years after Durack's investigation, his prediction regarding the publishing format came true – the last printed version appeared in 2004, and it is currently only available in on-line digital format. Similarly, in 2012 Encyclopaedia Britannica (first published in 1771) announced that its 2010 printing had been its last physical manifestation.

[2] This is probably a paraphrase of a statement by Samuel Johnson, as recounted in his 1791 biography by James Boswell [20].

While attempts at implementing artificial intelligence algorithms for medical decision support is almost as old as the development of digital computers themselves, Miller [22] points out that "*diagnosis is more than the act of associating the name of a disease or syndrome with the findings in a patient case*", and he cautions against a brute-force approach of "*mindlessly eliciting all possible patient data*" [23], as the implications of this would be "staggering" in terms of effort, cost, time and risk (mainly to the patient). Furthermore, diagnostic inputs as used by flesh-and-blood physicians are not just a list of "findings", but very often usefully include temporal data – i.e. information about changes in symptoms, over unstipulated time periods, possibly untreated or due to preliminary or exploratory therapy. Diagnosis itself is therefore also a process (not just a decision), which, due to the essentially unlimited variability in the input and output spaces, cannot easily be formularized, and thereby stubbornly defies optimisation. Nonetheless (or maybe specifically because of the challenges posed by this complexity), artificial intelligence research has been using medical data sets from the very beginning, with the first publications on the subject already appearing in the 1950s [24], [25].

Owing to the fact that the data used in medical decision-making processes are normally noisy, quite heterogeneous, and often incomplete, data fusion algorithms able to tolerate such variety and uncertainty may assist in improving the decision-making process. Although data fusion techniques have historically been developed primarily for military use [26], applications are also found in most aspects of civilian life. Since 1998 there has been an annual "International Conference on Information Fusion", organized by the International Society of Information Fusion under the aegis of the IEEE. Data fusion entails the combination of information from separate sources (on any level from raw data blending to decision synthesis [27]), normally with the aim of achieving a result that is in some sense better than can be achieved when simply using single-source data. "Better" can refer to aspects such as improved accuracy, usability, dependability or completeness, and may also include emergent results: for example, stereoscopic combination of data from spatially separated image sensors can produce range estimates – something which the individual sensors do not supply on their own.

### 2.2.2 Data fusion algorithms

#### 2.2.2.1 Symbolic learning – expert systems and heuristic algorithms

The most immediate and obvious approach to mechanizing the expertise of a knowledgeable practitioner in a given field is to attempt to capture this proficiency in the form of a set of rules that would emulate the decision-making ability of such a person. Decision guidelines of the IF-THEN-ELSE form are eminently suitable for computer implementation, but unfortunately experts are rarely able to describe their decision-making processes sufficiently succinctly to easily allow the required distillation into an unambiguously programmable set of rules. Often a physician would have a 'feeling' about a patient, which, according to Ledley and Lusted [24], is almost certainly based on complex (but often subconscious) reasoning processes, including assessments of the patient's appearance, expression, reliability, and history, as well as the more obvious and quantifiable results of medical examinations and tests.

Two main building blocks are required to construct an expert system: the *knowledge base* and an *inference engine*, with the latter 'reasoning' about the former, the way a human supposedly does. Some of the very first arguably successful artificial intelligence software projects were expert systems applied to medical subjects: in the 1960s the Stanford Heuristic Programming Project created the DENDRAL project, aimed at studying hypothesis formation and scientific discovery (targeting the identification of organic molecules based on chemistry knowledge and mass spectra), as well as MYCIN, an expert system built by Edward Shortliffe to identify infectious bacteria and recommend suitable antibiotic therapy [28], [29].

These early programs ran on time-shared mainframe computers, and were coded in Lisp, a general-purpose coding language created in 1958, with a strong symbolic computation orientation. Prolog (created in 1972), with its built-in inference engine, later replaced Lisp as the computer language of choice for artificial intelligence applications.

#### 2.2.2.2 Statistical learning

##### *2.2.2.2.1 Bayes' theorem*

In 1763 a paper was published in the *Philosophical Transactions of the Royal Society of London* that was to revolutionise statistics and probability theory several centuries later. The Reverend Thomas Bayes's posthumous paper titled "*An Essay Towards Solving a Problem in the Doctrine of Chances*" [30][3] contained a formulation of what has become known as Bayes' Theorem.[4] Amongst others, this theorem can be used for inference, by supplying a rational foundation for updating a subjective degree of belief to account for evidence. According to Sir Harold Jeffreys, Bayes' Theorem "*is to the theory of probability what Pythagoras's theorem is to geometry*" [32].

In its most common form Bayes' theorem as applied to two possible outcomes *A* and *B* is given by:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.1}$$

where *P*(*A*) denotes the probability of an outcome *A* and *P*(*A*|*B*) refers to the conditional probability of *A* given *B*.

Also known as the epistemological interpretation, Bayesian probability indicates the progressive updating of belief in a proposition while evidence accrues. For example, a naïve person, when shown a cubical die, may initially assume that there is a 50% chance of throwing a given number, say a *one*, in six tries. However, if a trial is done, the actual outcome can be used to update the expectation or belief, and, if the die happens to be fair, the expectation should eventually, after a large number of trials, settle at about 66.51%. (This

[3] edited and published by his friend Richard Price

[4] The possessive form for the singular noun Bayes may have been expected to be *Bayes's* – and indeed this usage predominated in the 18th and 19th centuries. However, according to Google NGrams (http://books.google.com/ngrams) the form *Bayes'* has supplanted it in print since the middle of the 20th century, and hence this form is used here, for the sake of consistency [31].

is because the probability of *not* throwing a specific number in six consecutive attempts equals $(\frac{5}{6})^6 \approx 33.49\%$, which is coincidentally close to, but not exactly, one third.)

Application of Bayes' theorem may also lead to some non-intuitive conclusions. As an example, take a hypothetical medical test, with a false positive rate of 1%, and also a false negative rate of 1%. In other words, an individual with the disease would, if tested, have a 99% probability of a correct, positive result ($P(A \mid B) = 0.99$), while a healthy person would have a 1% probability of a positive (incorrect) result. If, however, the incidence of the disease itself is low (affecting say 1 in 2500 people, or 0.04%, which is approximately the situation with CF in the United Kingdom [33]), then $P(A)= 0.0004$, and Equation (2.1) results in the possibly startling result that a positive test result only implies a 3.81% probability that the specific individual actually has the disease, despite the "99% accuracy" of the test. To understand this result it is mainly necessary to realize that a false positive rate of 1% will result in ten thousand errors out of a million tests, while actually only 400 (1 in 2500) of the same million people are expected to actually have the disease. Additionally the 1% false negative rate would lead to an average of 4 of those affected individuals erroneously being declared healthy.

#### *2.2.2.2.2 Dempster-Shafer theory (belief networks)*

The Bayesian approach requires that the relevant probabilities (either modelled or statistically determined) for each contributing factor be known; from this the resultant probabilities are computed as above – this is also known as the *naïve* Bayes approach. In 1968 Arthur P. Dempster published a paper titled "*A Generalization of Bayesian Inference*" [34], which, together with the work of Glenn Shafer some years later [35], [36], has become known as the Dempster-Shafer theory.

Dempster-Shafer theory uses what is termed a *degree of belief*, called a *belief function,* instead of the more conventional Bayesian probability distribution. A belief function assigns probability values to sets of possibilities (not necessarily single events). These belief functions use the probabilities of a related question to derive a degree of confidence (or belief) for the actual question.

When used for sensor (or data) fusion, this approach entails the determination of subjective probabilities for a related question, from which the degree of belief for the actual question is derived, as well as Dempster’s rule [34] which governs the combination of such degrees of belief, by reflecting the general assumptions about the data. This includes whether the degrees of belief are actually constructed from independent evidential information. Such combination can even include beliefs based on hints [37], opinions, or preferences [38].

As an example, suppose we have a friend called Alice, about whom we have subjective opinions regarding reliability: we think there is a 90% probability that she is reliable, and a 10% probability that she is unreliable. If she now tells us that the City Hall burned down yesterday, this testimony justifies a degree of belief of 0.9 that this had indeed happened, but no (zero) degree of belief that the City Hall had *not* burned down. This zero does not imply that we are quite sure that there had not been a fire at City Hall (as a *probability* of zero would have us conclude); it simply reflects the fact that her statement does not give us any reason to believe that there had *not* been a fire. This is because, if she is reliable, her testimony is per definition correct, but, if she is unreliable, it is not automatically untruthful. These two numbers (0.9 and 0) constitute our belief function.

Dempster’s combination rule can be demonstrated by introducing a new witness – if our friend Ben (who, like Alice, happens to have a subjectively assessed 0.9 probability of reliability and 0.1 of unreliability) also claims that the City Hall burned down, we can multiply these (independent) probabilities.

| | |
|---|---|
| *Both are reliable :* | 0.9 x 0.9 = 0.81 |
| *Both are unreliable :* | 0.1 x 0.1 = 0.01 |
| *At least one is reliable :* | 1- 0.01 = 0.99 |

Because they said the same thing we can therefore assign a belief of 0.99 to the event as claimed. This approach actually predates that of Bayes, with George Hooper already publishing monographs on the subject at the end of the 17th century [39].

Now, if they *contradict* one another (Ben says there was no fire), they cannot both be right, which means that at least one of them is unreliable – and possibly both are. The prior probabilities are:

| | |
|---|---|
| *Only Alice is reliable :* | 0.9 x 0.1 = 0.09 |
| *Only Ben is reliable :* | 0.09 |
| *Neither one is reliable :* | 0.01 |

The posterior probabilities (given that we now know that they cannot both be reliable) are:

| | |
|---|---|
| *Only Alice is reliable :* | $\frac{0.09}{0.09+0.09+0.01} = \frac{9}{19}$ |
| *Only Ben is reliable :* | $\frac{9}{19}$ |
| *Neither one is reliable :* | $\frac{1}{19}$ |

We therefore have a $\frac{9}{19} = 0.474$ degree of belief that the City Hall burned down (as Alice says), and the same degree of belief that it did not (as Ben claims). Note also that there remains an element of ignorance – Dempster's rule does not require the provision of prior probabilities that sum to one, such as in the traditional Bayesian approach.

The above demonstrates that we can obtain degrees of belief for one question (did the City Hall burn down?) from probabilities for a different question (in this case whether the witness is reliable). When we find conflict, the *a priori* assumption of independence (with respect to our subjective assumptions of probability) between the items of evidence is proven to be invalid, and has to be updated.

Although Dempster-Shafer theory is well suited to describe evidence, in a way which is compatible with the methods that humans use, it continues to elicit criticism, some of it from quite eminent scholars, such as the late Lotfi A. Zadeh [40] (also see the section on Fuzzy Set theory below), Judea Pearl [41], Andrew Gelman [42], and others [43]. These researchers claim that belief functions are often difficult to interpret, and that they are inadequate or inappropriate, especially when handling incomplete knowledge, because this can be demonstrated to lead to contradictions and incorrect decisions.

### 2.2.2.3 Neural networks and related algorithms

#### *2.2.2.3.1 Multilayer feedforward artificial neural networks*

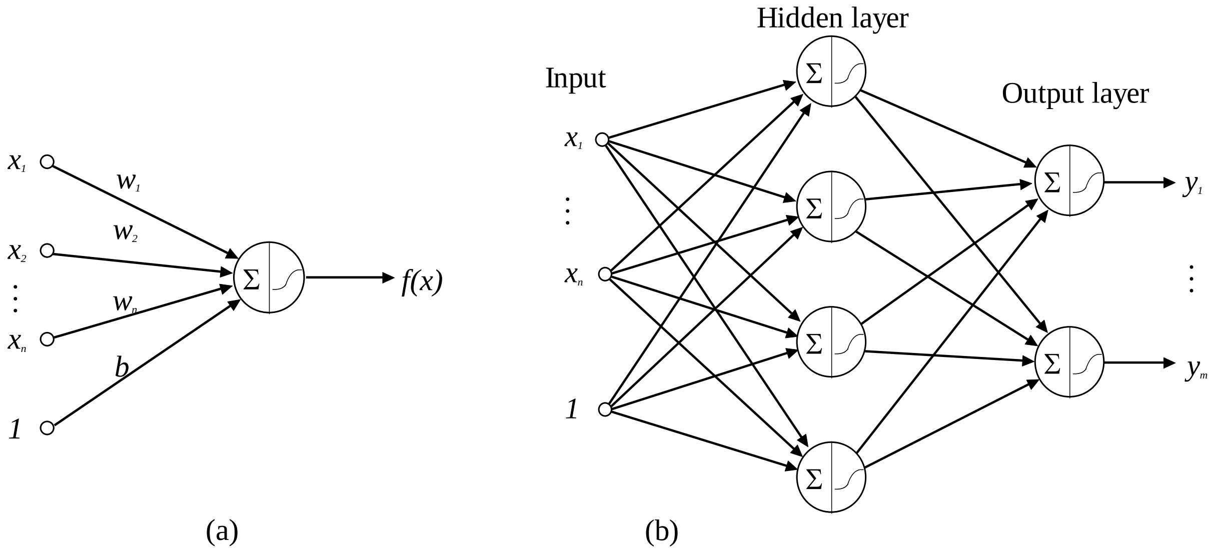


**Figure 2.1 –** (a) Perceptron (single neuron), (b) Fully interconnected multilayer feedforward neural net, with input nodes, one hidden layer, and output layer.

In 1958 Rosenblatt proposed the perceptron as a biologically inspired model with applications in data storage and machine learning [25]. The basic perceptron is a binary classifier, mapping an input vector *x* to a single binary output *f*(*x*) using a weight vector *w* and a bias (offset) value *b*:

$$f(x) = \begin{cases} 1 & \text{if } w \cdot x + b > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

However, the single-layered perceptron (also sometimes called a “neuron” – see Figure 2.1(a)) cannot solve non-linear problems (with the Boolean exclusive-OR function as a very simple problematic example), so in 1974 Paul Werbos introduced the backpropagation algorithm which could be used to “train” a multilayer network of neurons/perceptrons as in Figure 2.1(b), by a gradient-descent error-minimisation adjustment of the weight vector values. For this to work the Heaviside step function (or threshold) implicit in equation (2.2) is replaced by a bounded, continuously differentiable “activation function” to ensure that there is indeed a gradient to descend. This function is typically a sigmoid function such as the logistic curve $f(x) = 1/(1 + e^{-x})$, or sometimes the hyperbolic tangent function, both of which have positive derivatives everywhere [44], [45].

Other training techniques have also been proposed – Mazzoni *et al* [46] devised a biologically inspired reinforcement learning approach, which they based on measurements made on the visual cortex of monkey brains, reporting performance similar to that achieved by the backpropagation algorithm.

George V. Cybenko [47] then showed in 1989 that a three-layer network (containing a single “hidden layer” of sigmoidally activated neurons) between the input nodes (where the feature vector *x* is applied) and the output layer (which also consists of neurons, one per output of the network – see Figure 2.1(b)) is necessary and sufficient to approximate any arbitrary nonlinear function. While networks with two and even more hidden layers are sometimes used, this is not required, and training of such networks can be challenging.

Nonetheless, the availability of massive computing resources has opened the field of Deep Learning, which usually implement several layers, including one or more convolutional layers. A convolutional layer is not fully connected, but instead computes the cross-correlation of its input with a smaller kernel, as inspired by the work of Hubel and Wiesel on the visual systems of animals [48], [49] - for this work they shared one half of the 1981 Nobel Prize for Physiology and Medicine. In 1980 Fukushima introduced the “neocognitron”, consisting of convolutional and downsampling (averaging) layers [50], followed by Weng’s cresceptron [51], which uses max-pooling (finding the maximum value instead of the average of a region).

Training a neural network requires a set of tagged (known correctly classified) data which sufficiently represents the variability of the data features in the expected application. These examples are applied to the network, the actual output is compared with the desired value, and the training algorithm then adjusts the interconnection weights *w* to reduce the error. This process is run iteratively until some halting condition is reached.

The main challenge in designing and training a useful neural network is to ensure that it generalizes properly – this means that it will correctly handle data examples that it has not been presented with during training, which is the whole point of creating a classifier in the first place. Failure to achieve this is usually due to insufficiently representative training data, or to *overtraining*, which refers to the situation where the network has “memorized” specific examples in the training data, rather than extracting the similarities between examples of the same class. An effective way to avoid this phenomenon is to limit the complexity of the network (which usually means limiting the number of neurons in the hidden layer) to the lowest value which still facilitates the desired or maximal performance. By dividing the data into different parts used for training and validation, the point during the training process where performance starts declining may be identified, and the process can be halted, at which stage the best solution is used, or, preferably, the complexity is reduced (by eliminating neurons or connections), and the entire process is restarted, until an optimum is achieved.

Training of a neural network can be very computationally intensive, but this is in general no longer a problem (although it used to be a constraint in years past), due to the ever-increasing availability of processing power and storage space. Even though training can still consume significant amounts of time (especially when large feature spaces and data sets are involved), this happens off-line; the resultant feedforward network is computationally very efficient due to the fact that only a relatively small number of multiplications and additions (one each per interconnection weight) have to be executed, as well as determination of the activation function value for each neuron in the hidden layer (which is usually done very efficiently via a lookup table).

Neural networks can handle noisy or incomplete data fairly well (due to the generalization characteristic), as long as sufficiently representative training information has been used. It does not, however, offer a clear exposition as to why a specific output is chosen [52] – although careful inspection of the interconnection weight values and the resultant hidden layer outputs may offer some insight, the internal representations of information are usually quite abstract, and offer little in the way of elucidation as to why a specific output ensued. Neural networks are therefore very much a "black box" type of classifier.

###### *2.2.2.3.2 Support vector machines*

Support vector machines (or networks) are similar to feedforward neural networks, in that they attempt to separate two classes of data by constructing a hyperplane (or set of hyperplanes). Originally invented as a linear classifier by Vladimir Vapnik in 1963, the current standard soft margin formulation (which allows for mislabelled examples) was proposed in 1995 by Cortes and Vapnik [53]. Additionally the so-called "kernel trick" uses a nonlinear function to transform the feature space. This allows a maximum-margin separating hyperplane which, if transformed back into the original input, may be highly nonlinear. This is very similar to what happens in a three-layer feedforward neural network as described above.

Support vector machines have been reported to perform well in various medical diagnostic applications [54], [55], [56], [57], after supervised training on suitably prepared data sets.

Although they are computationally comparatively cheap to train, support vector machines suffer from the same limitation as the feedforward neural network, in that the parameters are not easily interpreted, meaning that results (outputs or decisions) are not really explained – it too acts as a “black box” classifier. Furthermore it can only be applied to separation of two classes at a time, necessitating the reduction of multi-class problems into a series (or tree) of binary decisions.

#### *2.2.2.3.3 K-Nearest neighbour algorithm*

The *k*-nearest neighbour algorithm is conceptually a very simple machine learning approach. It defers computation until a classification is required; at that point the training set is searched to find the *k* nearest examples (where *k* is a positive integer, usually small and odd). If $k>1$, a majority vote is taken (thus the preference for an odd value of *k*) to determine to which class the new example should be assigned.[5]

While the *k*-nearest neighbour algorithm is simple to implement, it is sensitive to noise, to the scaling of features relative to their importance, as well as to irrelevant features. There are techniques to optimise feature scaling and also the value of *k* (for example by using evolutionary algorithms [58]), but execution can be computationally demanding, especially in the case of a large training set (which is desirable from an error-rate point of view), as the training set has to be searched each time to identify the nearest neighbours.

Owing to the fact that the decision is explicitly based on specific known examples, this algorithm does have the advantage that it can offer some justification as to why it gave the answer it did – the user can easily be referred to similar examples in the training data (one or more of the just-identified *k* nearest neighbours with the same classification as the eventual answer). This is often perceived as satisfactory by users [52].

[5] A similar approach is sometimes used for regression, where the suitably weighted average of the *k* nearest points is assigned to the new value – in fact this is a generalization of linear interpolation.

#### *2.2.2.3.4 Fuzzy set theory*

In 1965 Lotfi A. Zadeh introduced the concept of fuzzy sets [59], which generalises the classical idea of sets to allow set elements various degrees of membership, described by a *membership function* ranging over the real unit interval [0,1].[6] Fuzzy set theory refers to conventional bivalent sets (where an element either belongs, or does not belong) as *crisp sets*.

The definition of a fuzzy set takes the form of a pair $(U, m)$ with $U$ being a set, and the membership function $m: U \rightarrow [0,1]$, where each $x \in U$ has a grade of membership $m(x)$. For an element $x \in U$ with $m(x) = 0$, the element is termed *not included*, while all other values constitute what is known as the *support* of $(U, m)$ – symbolically this is the set $\{x \in U | m(x) > 0\}$. The *kernel* of $(U, m)$ consists of the *fully included elements*, constituting the set $\{x \in U | m(x) = 1\}$.

Fuzzy theory is eminently suitable for applications where noisy or incomplete information is to be processed – the first publications into the use of fuzzy sets for medical decision support were by Hiramatsu *et al* [61] in 1974 and Wechsler [62] in 1976.

Although fuzzy set theory facilitates manipulation of data with real-valued membership functions as defined above, the ambition is usually to arrive at some definite result, such as a decision, an action, or a diagnosis. Because an element can be a member of several sets to various degrees, techniques have been developed to *defuzzify* the result. The simplest approach would be to choose the highest membership value, in the process discarding all other contributions – but such a loss of information is rarely a good idea.

[6] A further generalisation of this idea was introduced shortly afterwards by Joseph Goguen, a student of Zadeh, who extended the idea to intervals other than the unit interval [0,1] – this is known as *L-fuzzy sets* [60].

A plethora of defuzzification techniques have been developed to address this issue, and, as usual, the best method depends on the application. Van Leekwijck and Kerre [63] conclude that the so-called *maxima* methods fare well in fuzzy reasoning applications, while *distribution* and *area* methods are suitable for use in fuzzy controllers.

### 2.2.3 History of computer-based medical diagnostic or decision support

The 1959 *Science* article by Ledley and Lusted titled "*Reasoning Foundations of Medical Diagnosis*" [24] identified most of the methodologies that would be followed subsequently by AI researchers attempting to improve medical diagnosis and risk assessment. They also anticipated most of the challenges that researchers and medical practitioners would encounter. In this paper they discuss symbolic reasoning (as used in expert systems), Bayes inference and conditional probabilities, heuristic reasoning, and, very importantly, they apply Von Neumann game theory principles to derive a value/risk assessment as additional input to the decision-making process. They furthermore realized that even just a few hundred diseases and a similar number of findings result in ridiculously large combinatorial spaces, which have to be 'reduced' to manageable levels – for this they proposed an ingenious system of notched filing cards, but clearly they were also aware of the growing possibilities of the nascent digital computer.

Shortly afterwards (in 1960-1961) the first operational Bayesian computer-based diagnostic decision support system was created by Homer R. Warner, for congenital heart disease diagnosis at LDS Hospital in Salt Lake City [64]. A number of important lessons were learned in this project. Warner realized that assumptions regarding the independence of diseases and symptoms were needed to simplify the Bayes calculations, and to this end he developed techniques to eliminate redundant (non-independent) data from case findings. He derived the probabilities required for the Bayesian calculations from his own case findings, as well as from literature. He also noted that his system was quite sensitive to database errors (presumably due to data entering mistakes) as well as to false positive findings, and he identified the need for an independent "gold standard" by which to judge the performance of his diagnostic support system.

Warner's approach required the user to enter all known 'findings', after which the computer would run through the calculations and produce a differential diagnosis. While this could work for a narrowly specialized area, general diagnosis involves too many findings and possible diseases to make this practical – this is where heuristic techniques as introduced by Gorry and Barnett in 1968 are of value [65], [66]. Their approach uses heuristics[7] derived from Bayesian analysis – a serial questioning strategy is used to obtain information, resulting in a sequential diagnosis model, which takes into account the cost of tests and errors. Heuristic programming is used to discard unlikely or irrelevant conclusions – as such it also takes into account observed statistics and dependencies, and uses this to reduce the problem space to tractable levels. This was also the approach used in the DENDRAL project, started by Joshua Lederberg in the mid-1960s [67]. This project is widely accepted as the first successful "*expert system for scientific hypothesis formation*".

Edward Shortliffe's MYCIN (which was based on DENDRAL) eventually had a knowledge base of about 600 rules and a relatively simple inference engine [28], [29]. It interrogated the operator (mainly via a series of yes/no questions), and issued a list of possible diagnoses (in this case bacterial infections), in descending order of confidence. Significantly, it would also recommend a suitable therapy, and supply its "reasoning", by listing the questions and rules that significantly contributed to the eventual diagnostic decisions and therapeutic suggestions. MYCIN used "certainty factors", because Shortliffe and his co-workers initially argued that classical Bayesian statistics would necessitate unrealistic independence assumptions, or the determination of an unfeasible amount of conditional probabilities. However, it was later shown that the certainty factor model also contains implicit assumptions, and that a probabilistic interpretation was indeed possible [68].

[7] Heuristic, from the ancient Greek 'eureka' (εὕρηκα), meaning "I have found", here denotes an empirical rule of thumb.

MYCIN was found to fare better than infectious disease experts tested on the same criteria. It proposed an acceptable therapy in 69% of cases, but it took 30 minutes or more to enter the data, and as a result it was never used in practice.

The Quick Medical Reference (QMR) project, created by Miller *et al* [69] in 1986 eventually grew to include approximately 600 “significant diseases” associated with about 4000 findings, using statistical data as well as an expert knowledge rule base. Reformulated in 1991 as a probabilistic model (called QMR-DT, for QMR Decision Theoretic) by Shwe and Cooper [70], the diagnostic challenge is to start with a subset of findings, and use this to infer a disease probability distribution.

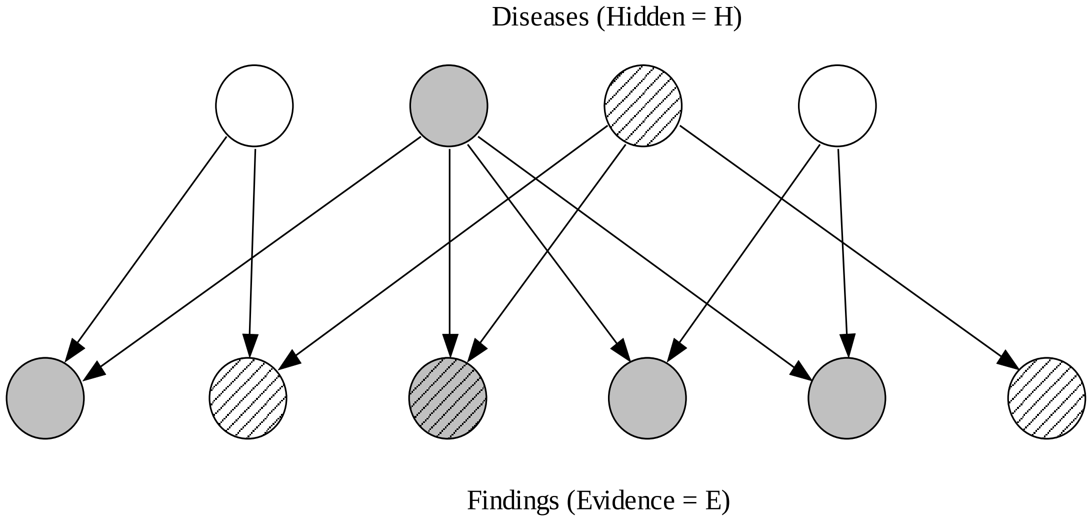


**Figure 2.2** – QMR-DT as a densely connected graphical network. With observed evidence or symptoms (which can be present, absent or unknown), the posterior probability of the diseases (hidden nodes) are to be inferred [71]**.**

Since the late 1960s there have been a huge number of medical diagnostic support projects, and right from the start researchers seem to have felt a compulsion to name their projects. This may possibly reflect a hope that it would eventually grow into a viable tool and potentially a commercial product or it may simply be because the most visible result of the research effort invariably is a computer program, which has to be named to enable the computer operating system to distinguish it from other programs and files. The following table is not comprehensive, but it does list a number of the most significant named efforts:

**Table 2.1** – Notable named diagnostic support projects.

| Title | Researchers/Creators | Type | Year |
|---|---|---|---|
| CONSIDER | Lindberg, Rowland *et al* [72] | Expert System | 1968 |
| MYCIN | Shortliffe [28], [29] | Expert System | 1973 |
| PIP | Pauker, Gorry *et al* [73] | Expert System | 1976 |
| CASNET | Weiss, Kulikowski *et al* [74], [75] | Bayes, *k*-NN | 1978 |
| RECONSIDER | Blois, Tuttle, Sherertz [76], [77] | Expert System | 1981 |
| CADIAG-2 | Adlassnig & Kolarz [78] | Fuzzy | 1982 |
| Internist-I/ Caduceus | Miller, Pople & Myers [79] | Expert System | 1982 |
| AI/COAG & AI/Rheum | Gaston, Lindberg *et al* [80] | Expert System | 1983 |
| SEEK | Politakis & Weiss [81] | Expert System | 1984 |
| Pathfinder | Horvitz, Heckerman *et al* [82] | Expert System | 1984 |
| QMR | Miller, Masarie & Myers [69] | Heuristic/Expert | 1986 |
| DXPlain | Barnett, Cimino *et al* [83] | Expert System | 1987 |
| ILIAD | Warner *et al* [84] | Bayes/Expert | 1988 |
| DESKNET | Yoon *et al* [85] | Neural Net | 1988 |
| Meditel | Waxman & Worley [86] | Probabilistic/Bayes | 1989 |
| TraumAID | Clarke, Niv *et al* [87] | Expert System | 1989 |
| QMR-DT | Shwe & Cooper [70] | Probabilistic/Bayes | 1991 |
| WebMD | Clark & Nigam | Unknown | 1996 |
| Wizorder | Geissbuhler, Miller [88] | Expert System | 1998 |
| NeuroShell | Kanagaratnam *et al* [89] | Neural Net | 1999 |
| Renoir & Pneumon-IA | Godo *et al* [90] | Fuzzy | 2001 |
| PROCFTN | Belacel & Boulassel [91] | Fuzzy | 2004 |
| Promedas | Wemmenhove, Mooij *et al* [92] | Bayes | 2007 |
| Mediquery | Carvalho, Isola *et al* [93] | Neural Net/Bayes | 2011 |
| Symcat | Monsen & Do | Bayes | 2011 |
| IBM Watson | Ferrucci *et al* [94], [95] | Expert/Bayes | 2013 |

Graphical models are tools which facilitate the representation and analysis of the relationship between arbitrary numbers of complexly-linked random variables [96], [97], [98], which have found application in a variety of statistical applications. The QMR-DT network as shown in Figure 2.2 consists of a simple graph – on the top level there are approximately 600 diseases, and on the bottom there are about 4000 potential findings, which are the observations that should be used to infer the probability of the (unknown) diseases. The problem is that the graph is densely connected – each finding or symptom can be associated with many diseases, and vice versa, which results in an inference problem that is NP-hard [71]: there are roughly $2^{600}$ potential disease hypotheses [99]. This clearly rules out any hope of exact inference using brute-force methods on current computer hardware.

A number of approximation techniques have been developed to address this problem – where exact inference is impractical the main approaches are sampling algorithms (mostly *Markov Chain Monte Carlo* and *importance sampling*) and so-called *variational algorithms* [97]. Additionally hybrid approaches are sometimes used, in which local optimisation using exact inference is applied in a global sampling or variational framework [100].

In the mid-1990s the first web-based self-diagnosis tools started appearing (e.g. WebMD, and later MSN Health and Fitness[8]), and with them a new psychiatric condition called *cyberchondria*, which has been described as "*the unfounded escalation of concerns about common symptomatology, based on the review of search results and literature on the Web*" [101]. This is hardly surprising, given that if even medical professionals cannot hope to fully grasp all possible symptom-disease association probabilities, the layman who suddenly has essentially unfiltered access to what used to be esoteric information is quite defenceless. [9]

[8] Respectively at http://www.webmd.com and http://healthyliving.msn.com

[9] The underlying condition of course long predates the internet – in his 1889 book *Three Men in a Boat* Jerome K. Jerome relates how he supposedly diagnosed himself (from a book) to be suffering from "*every other known malady in the pharmacology*" except housemaid's knee. His doctor's terse advice: "*… don't stuff up your head with things you don't understand.*"[102]

Furthermore there have been allegations of conflicts of interest, with web-based medical advice sites being investigated for preferentially advising the use of products and services from partnered pharmaceutical companies, with little if any medical motivation. Such “medical” advice should therefore be taken with a healthy dose of scepticism, or not at all.

### 2.2.4 Current status

Currently IBM’s Watson[10] project is by far the most visible contender in the race to create a practical tool that will improve the quality of medical diagnoses and treatment decisions. After defeating the reigning world chess champion Gary Kasparov with their *Deep Blue* chess computer, IBM created a new project in 2006, which they named *Deep QA*. As their primary target they selected the American television quiz program *Jeopardy!* which required that their machine should be able to process natural language, generate hypotheses, and to improve using evidence-based learning [95]. In February 2011 they achieved the intended spectacular success, when Watson famously defeated two former *Jeopardy!* champions in a televised match, and later that month also overwhelmed five members of the United States House of Representatives in a similar quiz match.

Immediately after this highly publicized showcase IBM announced that they would be applying the Watson technology to a variety of more practical applications, with health care at the top of the list [94], [103]. According to IBM press releases they are grooming Watson to take the US Medical Licensing Examination, although they are very careful to stress that it is simply a support tool to medical practitioners. Its natural language capability enables Watson to absorb massive amounts of *unstructured data* (information not specifically formatted for machines, such as medical research publications, most patient medical records, physicians’ notes, and historical medical statistics) and to use that to generate diagnoses and treatment suggestions which potentially take into account vastly more current medical knowledge than any human practitioner could hope to encompass. In a cooperative

[10] Named after IBM’s erstwhile CEO and chairman Thomas J. Watson.

agreement with the Memorial Sloan-Kettering Cancer Center, IBM is currently focusing on improving lung and breast cancer therapy [103].

### 2.2.5 Critical discussion

The general move to digitize all information has created a new phenomenon known as Big Data (with “big” currently considered to be terabytes up to exabytes in size), that refers to the massive on-line data sets which have lately become available. These data sets are characterized by the “three V’s” (variety, volume and velocity) [104], [105] and have become the focus of many commercial data mining efforts, including finance and economics, social networking, meteorology, and medical decision support. Many Big Data efforts are commercial in nature, and therefore the developers of these systems – mostly large companies such as IBM and Cray (via their Yarcdata subsidiary) – tend to keep the algorithmic implementation details proprietary.

In addition to facilitating better diagnosis and treatment, data mining and modelling may also offer the potential of uncovering hidden dependencies between factors such as diseases, symptoms, living conditions, genes, and even health policy. This should be of great interest to clinical researchers, health officials, and possibly even medical aid administrators.

Medical decision support tools have repeatedly been shown to outperform clinicians [29], [106], [107], [108], yet only a selected few applications have gained a foothold in medical practice, in narrowly focused applications such as interpretation of electrocardiograms, arterial blood gas data or pulmonary function tests. A variety of reasons exist for this relatively low penetration, despite more than half a century of research – several of these are addressed in the following subsections.

#### 2.2.5.1 User resistance

Understandably some physicians feel intimidated or threatened by decision support systems; this may potentially hinder acceptance of such tools. The IBM publicity machine is careful to stress that Watson is merely a tool being developed “to improve the quality of health care”. More than sixty years ago Ledley and Lusted [24] cautioned that their proposed approach “*in no way implies that a computer can take over the physician’s duties*”, but that the task of the physician is actually likely to become more complex; the pay-off would hopefully be better diagnoses and “*a more scientific determination of the treatment plan*”. This indeed seems to have happened as predicted.

On the other hand, in informal interviews with the author, medical practitioners indicated that most of them decry the unfettered access that patients have to medical information, or at least the indiscriminate acceptance of any and all information thrown up by an internet search or web-based diagnostic tool. Increasingly they report the aggravation of having to explain and motivate each of their diagnoses and decisions to cyberchondriacs who lack the required background to understand them, and who far too often direct their distrust, scepticism and even aggression at their (supposedly qualified and experienced) doctor or pharmacist, rather than at Wikipedia or WebMD.

#### 2.2.5.2 The knowledge acquisition bottleneck

Shortliffe’s early work on MYCIN highlighted the difficulty of extracting the knowledge base from human experts - the ‘knowledge acquisition bottleneck’ [29]. This remains a huge challenge – Miller estimates that more than 40 person-years had been expended between 1973 and 1999 on the knowledge base for INTERNIST-I, later superseded by QMR [22], yet both these ventures have been abandoned as active medical decision-support projects.

#### 2.2.5.3 User interface

Another stumbling block is the time and effort required to answer questions and enter data when handling an actual patient case. Very early on Shortliffe realized that, although his MYCIN expert system could arguably outperform humans (at least in its limited field of expertise), it remained impractical, due to its cumbersome and time-consuming user interface. Only recently, with digital patient records becoming fairly ubiquitous at most medical practices, has this limitation started to reduce in significance. Nonetheless, data are still rarely structured for machine understanding, which is why the IBM Watson approach most likely represents the rational way forward: accepting that humans mostly communicate using natural language, and focusing on understanding that. Hitherto, the only way to use decision support tools was for medical personnel to effectively learn a new and restricting language. IBM aims to circumvent this limitation.

#### 2.2.5.4 Computational limits

The immense challenge posed by the general diagnosis and treatment problem as exemplified by the QMR network [69], [99] is undoubtedly a large part of the reason why IBM chose to limit the field for their first real-world applications to selected cancer sub-specialties. Big Data currently being mined by meteorologists, scientists, economists and others require massive infrastructure in terms of storage and processing power, which limits the field to large players with deep pockets, such as governments, IBM, Microsoft, Cray, Amazon, Facebook, and Google (which is probably the biggest of them all, with an estimated 10 exabytes of data storage capacity spread over at least 13 server farms worldwide).[11]

#### 2.2.5.5 Legal and ethical considerations

Finally there also remain the ethical and legal questions:

- Who is responsible for a wrong diagnosis or recommendation of incorrect or sub-optimal therapy? Here it is not foreseen that there will soon be a shift from responsibility away from the clinician. While getting medical decision support systems to pass a medical licensing examination will be a Turing Test *tour de force* of artificial intelligence [109], the general view is that these systems will for the time being remain merely tools, like stethoscopes and magnetic resonance imagers [23], [110]. The real risk lies in that medical practitioners may become over-reliant on such tools, possibly neglecting to apply the full focus of their attention and experience to the patient.

[11] See http://www.google.com/about/datacenters/inside/locations/index.html. An exabyte is $10^{18}$ bytes. The volume of all the printed text material in the US Library of Congress is estimated at about 10 terabytes ($10^{13}$ bytes), which means Google can store it a million times over.

- How can privacy and confidentiality be guaranteed? With huge data breaches regularly in the news, users are understandably wary of entering potentially sensitive personal data into a system.
- Who owns the data entered into the decision support system? No doubt the large players will attempt to monetise their considerable investments into decision support system development, but this immediately raises the spectre of conflicts of interest.
- How can institutional bias be detected and avoided? The great disparity between medical care and research spending in different parts of the world has resulted in differences in knowledge and diagnostic focus which can easily manifest as discriminatory [8].

## 2.3 ESTABLISHMENT AND EQUILIBRIUM LEVELS OF DELETERIOUS MUTATIONS IN LARGE POPULATIONS

The focus is now shifted from inferring the presence of a variation in an individual, by amalgamating known symptoms, to deducing the magnitude of possibly unknown pathological processes due to genetic variations in entire populations, by exploiting census data.

### 2.3.1 Introduction

Heterozygous carriers of some common mutations, e.g. in the cystic fibrosis transmembrane conductance regulator (CFTR) gene, seem to have a survival/fecundity advantage compared to non-carriers [2], [111], [112], while the homozygous state results in a definite disadvantage. An example of a heterozygous advantage that is environmentally dependent is malaria resistance as a consequence of specific mutations in the haemoglobin gene [3], [4], bringing with it the risk of sickle-cell anaemia in homozygotes. Another common example of a heterozygous advantage is the major histocompatibility complex (MHC) in vertebrates [113].

The implications of such a benefit (termed ‘selective advantage’) were investigated by Haldane, who focused on the mathematical probability of purely beneficial mutations becoming established in a population of size $N$ [114]. Haldane’s results were subsequently refined by Wright, Kimura, and others [9], [115]. Wright also introduced the notion of ‘effective population size’ ($N_e$), to account for effects such as non-random mating, inbreeding and unequal sex ratios, which may influence the effectiveness of natural selection forces.

For human populations we further extend this approach by drawing on the results of Dunbar and Lehmann to propose a realistic range and upper limit to the size of the group from which an individual is likely to select a mate [13], [14]. This is analogous to the ‘breeding unit’ introduced by Wright in 1946, which he termed $N_\mathrm{n}$, being the spatially closest individuals in a circular area with a radius of $2\sigma$ [116]:

$$N_\mathrm{n} = 4\pi\sigma^2 d \tag{2.3}$$

with $\sigma$ being the standard deviation of a spatially distributed 2-dimensional normal distribution around an individual and *d* the areal distribution density (i.e. individuals per unit of 2D space). This reflects the observation that the parents of an individual organism are more likely to be proximate than remote, and, as Nunney suggests, that $N_\mathrm{n}$ will be relatively constant, subject to the assumptions that $\sigma$ is characteristic of the species and that there is an inverse relation between dispersal and density, i.e. that one can normalize for $d$ [10]. This leads to a parental probability distribution solely dependent on distance *r*:

$$f(r) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-r^2}{2\sigma^2}} \tag{2.4}$$

In the case of human populations, we propose that the concept of distance ($r$) should be reinterpreted as social proximity, rather than necessarily physical proximity, due to the global mobility (i.e. potentially high dispersal) that has been attained by humans in recent times, and which does not necessarily translate into an increase in $N_\mathrm{n}$ (which is also the pool from which an individual would usually select a mate).

In their work performed on the genetics of guinea pigs and fruit flies, Wright and Dobzhansky additionally had to introduce an inbreeding factor *F* and an immigration index *m* [116], [117], [118], [119], to compensate for the effects of self-fertilization, and the postulated (but never quantified) expectation that ‘immigration’ is likely to be local, from adjacent localities which will probably resemble the target locality (in a gene frequency sense) rather than that of the entire species; this immediately requires yet another estimated adjustment to obtain an *effective* value for *m*. All these complications are obviated in our neighbourhood definition for humans (in which self-fertilization can also be neglected), where breeding structure is abstracted to the social dimension, and inbreeding and genetic drift become emergent features, rather than arbitrary modifying requirements to force the model to fit actual observations.

Based on primate studies and the size of the human neocortex, Dunbar posits a nominal maximum group size of 148 for humans (usually rounded up to 150), with a 95% confidence interval between 100 and 230 [13], but also stresses that this is an upper value, which is only approached under extreme environmental stress, where the significant time and energy investments of maintaining close social bonds are repaid by the survival benefit realized by being part of a larger group. In more prosperous times, group sizes tend to reduce.

This aspect is one of the major novelties of the framework proposed in this thesis, which allows the simulation of large populations while modifying the community size ($N_n$) in order to create scenarios to study how local selection affects the pattern of deleterious and/or advantageous variants.

The effect of this community size is then explored, in conjunction with the selective advantage of a given mutation, on the probability that a mutation will become established in the population, and on the eventual equilibrium levels that are reached, especially in the case of mutations that are beneficial when heterozygous, but pathogenic (or less beneficial) when homozygous. Such mutations do not necessarily become ubiquitous once established, a possibility that Wright and Dobzhansky mistakenly dismissed in their seminal work on lethal mutations in fruit flies [120]. Conversely, if the occurrence levels of such a mutation in a population is known, the model can be used to estimate the selective advantage that it confers on heterozygous carriers, *without requiring any knowledge of the specific manifestation and mechanism of such a selective advantage.*

### 2.3.2 Establishment of mutations

For haploids, the selective advantage (also called the selection coefficient) conferred by a specific mutation is defined as an additive term $s$, such that carriers would, on average, have (1+$s$) times the wild-type (i.e. non-carrier) number of offspring [12]. Note that $s$, normally termed a ‘selective advantage’, can also be a negative number, implying a disadvantage by resulting in fewer offspring of mutant carriers, whether through lowered fertility, or decreased survival to procreative age (which in some sense is the same thing - irrespective of the mechanism, the result being a reduction in the number of offspring compared to wild-type individuals). Although the case of $s < 0$ has the physical interpretation of a selective *disadvantage* (also known as purifying selection), the case of $s < -1$ is meaningless, and hence the parameter $s$ should be constrained on the interval $[-1,\infty)$.

Using a deterministic model, any beneficial mutation (that is, with $s > 0$) will inevitably grow in prevalence, guaranteeing eventual fixation in the population. However, in reality genetic drift causes random fluctuations in the frequency of lineages, which can easily extinguish even highly beneficial mutations when their prevalence is low, as will be shown. A stochastic treatment is used to analyse such situations, which especially apply whenever a new mutation appears *de novo* in a single individual. The mutation will only be established in the population (and only then a deterministic model may be applicable) if this mutation survives genetic drift.

During admixture between different populations, ‘new’ alleles are introduced at significant levels into both groups. Under such conditions (relatively large populations and high prevalence) the alleles could be considered to already be established and therefore to be less subject to the vagaries of genetic drift, but rather with their eventual fate dominated by the relative selection coefficients that the alleles confer (i.e. closer to the deterministic case).

Addressing the fixation probability $P$ of a single copy of an advantageous allele in a large population, Haldane found that $P \approx 2s$ if $s$ is small [114]. Barrett *et al* [121], drawing on this as well as on the subsequent analyses for finite populations by Kimura [122], [123], show that

$$P \approx 1 - e^{-2s} \tag{2.5}$$

which also accurately approximates the probability that a single advantageous allele with a large positive value of $s$ will survive stochastic loss. Negative or zero values of $s$ always lead to eventual extinction. Migration within the population under consideration has no effect, as panmixia (where all individuals are equally likely to be mates, irrespective of distance) is implicitly assumed. This implies a homogeneous landscape, whether it be geographic or social, and therefore transplanting an allele does not impact its viability.

### 2.3.3 The price of success

#### 2.3.3.1 Homozygosity

A large number of deleterious recessive monogenic autosomal variations have been identified in humans. Many of these have attained wide penetrance, which suggests the presence of some heterozygous selective advantage.

**Table 2.2** – Some deleterious recessive monogenic autosomal variations.

| Condition | Possible heterozygous advantage |
|---|---|
| Cystic fibrosis | Increased resistance to cholera and typhus [111] |
| Sickle-cell disease | Increased resistance to malaria [3] |
| Thalassaemia | Increased resistance to malaria [124] |
| Tay-Sachs disease | Enhanced resistance to tuberculosis [2] |
| Gaucher’s disease | Enhanced resistance to tuberculosis [125] |
| Spinal muscular atrophy | Resistance to poliomyelitis [126] |
| Hereditary haemochromatosis (‘Celtic Curse’) | Protection against celiac disease [127], enhanced resistance to typhoid fever and tuberculosis [128] |
| Oculocutaneous albinism | Protection against tuberculosis and leprosy [129] |
| Nonsyndromic deafness | Thicker skin, resistance to infection [130] |
| Bardet-Biedl syndrome | Enhanced fat storage [131] |
| Phenylketonuria | Improved resistance to ochratoxin A, fewer mycotic abortions [2] |
| Friedreich’s ataxia | Increased iron accumulation [132] |
| Niemann-Pick Type C | Increased resistance to filoviruses (e.g. Ebola and Marburg) [2] |
| Congenital disorder of glycosylation | Increased resistance to glycosylation-dependent viral infections (e.g. influenza, hepatitis C, HIV-1) [133] |

A mutation that is purely beneficial will completely displace the wild-type allele only if it successfully runs the gauntlet of genetic drift while still rare. In a diploid population, however, heterozygous carriers may derive a selective advantage from a given mutation, while homozygosity results in a reduced selective advantage (or even disadvantage), such as in the case of CF, sickle-cell disease and others – see Table 2.2. As the prevalence of such a mutation in a population increases, the probability of producing homozygous offspring also rises, to the point where the relative disadvantage of homozygosity exactly balances the heterozygous advantage. An equilibrium is reached, and this depends on the relative magnitudes of the effects, as well as the population parameters, especially the effective population size $N_e$ and the neighbourhood size $N_n$.

### 2.3.3.2 Environmental factors

The striking geographic correlation between the distribution of the sickle-cell allele and the prevalence of malaria [3] demonstrates the importance of external factors on the selection coefficient of a given genetic variation. As long as the local population is exposed to malaria, the sickle-cell mutation (if present) confers an advantage ($s > 0$) to heterozygous carriers and rises to prevalence levels limited by the negative effects caused by the associated increase in homozygous individuals. Where malaria is absent, there is no selective advantage ($s$ may even be slightly negative), and the allele becomes extinct. Similar correlations between infectious diseases and genetic variations have been identified for many of the conditions listed in Table 2.2. This goes a long way towards explaining why such apparently deleterious alleles have not yet been eliminated through natural selection.

### 2.3.3.3 Migration, selection and inbreeding

A shortcoming of Haldane’s approach is that the population $N$ is assumed to be large, constant and with equal sex ratios and random mating (panmixia). This is not normally the case. Many subsequent researchers have addressed this [9], [11], [12] and have introduced the concept of an effective population size $N_e$ which would result in the same variance as the current population under consideration. Usually the effective population size is smaller than the census size ($N_e < N$), and there is the even smaller community size (or neighbourhood number) $N_n << N_e$ which affects genetic differentiation between subpopulations: the smaller $N_n$ the larger the differentiation between them, due to the decreased dispersal distance and increased genetic drift, considering limited or no gene flow [10].

### 2.3.4 Equilibrium levels

Surprisingly little has been published on the subject of the prevalence levels of deleterious mutations in diploid populations. Even Wright and Dobzhansky, after admitting to the possibility that there may be a ‘hypothetical class’ of lethal mutations (i.e. $s \rightarrow -1$ when homozygous) that benefit from $s > 0$ when heterozygous, erroneously conclude that “*such lethals should be detected easily”* [120], and, having failed to do that in their seminal fruit fly experiments [117], [118], [120], [134], [135], summarily dismissed this possibility. Interestingly, in the same era Dubinin [136] and others determined that many *Drosophila* mutations do indeed ‘*increase the viability of heterozygotes’.* However, even in the 1940s they seemed reluctant to directly contradict or criticize Dobzhansky, despite rather strong evidence. Dubinin merely acknowledges that their conclusions appear to differ.

# CHAPTER 3 DISEASE MODELLING AND ANALYSIS

## 3.1 INTRODUCTION

Diseases are complex, with many heterogeneous factors playing a role (as causes, results, or both). These features are also of different types, with differing types of uncertainty, and often there are intricate interactions between known (or postulated) contributing elements, as well as with other, unidentified factors; this vastly complicates efforts to analyse the disease. Through application of data fusion techniques to disease models, it is hoped to create a framework which can be used for study and improved understanding of the condition, its causes, and its temporal evolution. This in turn may facilitate improved personal medical care and better public health management and policy.

To achieve this goal, data will be gathered and used to build graphical Bayesian models. The data to be integrated may be primary measurements, higher-level outputs of probabilistic classifiers, or even derived from non-probabilistic sources such as human assessors. The generative property of Bayesian networks can then be exploited for “what-if” analyses, by conditioning on different possible outcomes and under different situations.

## 3.2 CASE-CONTROL STUDIES

Medical trials and studies take many forms, usually with the eventual goal of learning something about the interaction between diseases, causes and sometimes interventions (treatment). Trisha Greenhalgh proposes the following "hierarchy of evidence" ranking the relative significance of the different types of primary study which should be considered, especially when clinical intervention is planned [137]:

1. Meta-analyses and Systematic Reviews
2. Randomised controlled trials (with those producing definitive results – i.e. with non-overlapping confidence intervals – obviously outranking those which do not)
3. Prospective (cohort) studies[12]
4. Retrospective (case-control) studies
5. Cross-Sectional Surveys
6. Case Reports

Although case-control studies are ranked below randomized controlled trials and cohort studies in the hierarchy of evidence, they are usually faster and cheaper to conduct. Often, too, ethical considerations constrain the design of clinical trials, or even the use of data gathered during tests which can be considered to have been unethical [138], such as the Japanese biological warfare and vivisection experiments [139] and various German experiments [140] conducted during the Second World War, as well as several post-war American trials such as the Willowbrook study, where children with intellectual disability were deliberately infected with the hepatitis virus [141]. Medical researchers are therefore

[12] Cohort studies can also be retrospective – gathering and processing data on a population *post hoc*, rather than monitoring them forward over a (typically long) period of time. Although normally faster and cheaper to conduct than prospective cohort studies, retrospective cohort studies are exposed to significant risk of various types of bias, similar to that faced by case-control studies.

often limited to merely observing naturally-occurring events when studying the ætiology of disease.

Especially in the case of rare diseases, where cohort studies would require extremely large sample sizes and/or very long time windows, case-control studies offer the potential to relatively quickly establish a possible link between a disease and some putative cause or risk factor. Even though it is often argued that a correlation shown by a case-control study does not prove causation [142], a carefully conducted case-control study can strongly suggest such a link. An early triumph of this approach was the 1950 demonstration by Richard Doll and Bradford Hill of the link between smoking and carcinoma of the lung [143], which was subsequently vindicated by cohort studies; for example, it is now accepted that about 87% of lung cancer deaths in the United States can be attributed to tobacco smoking [144].

The modern era of case-control studies was arguably initiated by Jerome Cornfield, who showed in 1951 that the exposure odds ratio (dividing the number of subjects with the disease by the number without, or cases vs. controls) is the same as the disease odds ratio (the ratio between those exposed to the postulated risk factor and those not exposed), and that this also approximates the relative risk, on the assumption that the disease of interest is rare [145]. The requirement for this last assumption was later shown by Olli Miettinen to be “overly superficial and restrictive” [146]. Work performed by Harold Dorn [147] and especially by Nathan Mantel and William M. Haenszel [148] created the statistical tools that put retrospective studies on a firm footing, by addressing and compensating for confounding factors. In 1979 Philip Cole referred to Cornfield, Dorn, Mantel and Haenszel when he pointed out that the “giants” on whose shoulders epidemiologists have stood during the preceding 20 years when doing case-control studies, were all statisticians [149].

## 3.3 GRAPHICAL MODELS

### 3.3.1 Motivation

One of the problems with an observational study (whether prospective or retrospective) is that the gathered evidence only shows what happened under the circumstances that were present when the observations were made. Multiple factors interact to result in a specific outcome (such as a specific disease diagnosis) which we merely select for when we condition on a given variable; the circumstances associated with this outcome did not necessarily cause it, and the actual evidence says nothing about what would have happened under different conditions, which were not present – so-called "counterfactuals" [142], [150].

Furthermore, when multiple interacting potential causes and effects are involved, standard statistical approaches run into difficulties. Human diagnosticians, however, routinely and quickly integrate data of different types, from different sources, to make treatment decisions, albeit not always optimally.

What is needed is a way of integrating diverse data into a model, which can be validated against observations, and can then be manipulated to explore the expected behaviour when one or more parameters are changed; in other words, to make a causal prediction.

Graphical models as described by Judea Pearl [98], [150], and others [71], [96], visually represent factors (which can be diseases, symptoms, environmental parameters, treatments or any other factor deemed relevant) as nodes, with connections between them denoting conditional dependence assumptions, thereby encapsulating the joint probability distributions of the various factors. Building on the 250-year-old work of the Reverend Thomas Bayes [30] and that of the Russian mathematician Andrey Markov, graphical model theory can be and is used for probabilistic inference and causal prediction of three types: policy evaluation (the effects of interventions), counterfactuals, and mediation (queries regarding indirect and direct effects) [150].

### 3.3.2 Previous work and current status

The relatively new fields of "omics" (e.g. genomics, proteomics, and metabolomics) strive to unravel the complex interactions between genes and proteins or metabolites. The Mendelian inheritance model of dominant and recessive traits turns out to be not quite that simple, with epistasis resulting in massively intricate webs of genetic interaction [151], while most diseases additionally seem to depend on interactions between genes and environmental factors [152], [153], [154]. Graphical models, both Bayesian networks as well as Markov networks, are eminently suited to address this problem, and indeed the first publications on the subject appeared in the late 1990s.

Currently Jason H. Moore is one of the leading scholars on the subject of computational research into epistasis and gene-environment interactions. He uses a wide variety of machine-learning approaches, including cellular automata, genetic algorithms, ant colony optimisation, neural networks, random forests and especially multifactor dimensionality reduction [154], [155], [156], [157], applying them to the problems of determining a suitable model structure (as supported or suggested by the data itself), limiting the complexity thereof (to contain the curse of dimensionality as far as possible), and then to optimise the resultant graphical model, which can subsequently be used for interpretation.

Su *et al*, in their 2013 paper titled "*Using Bayesian networks to discover relations between genes, environment, and disease*" [153], use the expectation-maximisation algorithm both to optimise parameters as well as the network structure itself, using incomplete data. They caution, however, against over-enthusiastic application of a Bayesian network learned from (and consistent with) observed data for causal interpretation, due to the risk of there being unobserved variables which may influence the outcome. As early as 1996 Breslow identified graphical networks with latent (hidden/unobserved) variables as a challenging yet promising approach to the problem of causal interpretation [158].

There are few publications on studies which also attempt to take into account the temporal feature inherent in health care [159] – this seems to be a promising area for further exploration, as the longitudinal element inherent in cohort studies may possibly be incorporated by the time-dependence of dynamic Bayesian models (of which hidden Markov models are a subset [160]).

### 3.3.3 A graphical model for CF

#### 3.3.3.1 Purpose

A Bayesian network has been created as a framework to integrate different types of information associated with CF, and to explore some of the complex interactions between contributing and resultant factors. This can be achieved by exploiting the generative property of Bayesian networks to make causal predictions regarding interventions, counterfactuals and mediation.

The information to be used includes observations and measurements, for example symptoms and test results, but also information such as sociological data, family history, ethnicity, other disease diagnoses, and even unobserved (latent or confounding) postulated variables.

#### 3.3.3.2 Diagnostic and modelling inputs

Cystic fibrosis presents with a wide range of often non-specific signs and symptoms [1]. To explore and demonstrate the power of the BN, a number of inputs of various types have been selected. Because of the difficulty of obtaining observational data required for the joint probability distribution functions, not all known sources of evidence were added, but only a few examples of each type, for which representative values could be obtained or estimated:

#### *3.3.3.2.1 Symptoms*

- In infants and young children: recurrent respiratory symptoms; failure to thrive; pancreatic insufficiency (present in up to 90% of cases) leading to steatorrhea, diarrhoea, and abdominal distension.
- Older patients: recurrent respiratory symptoms (often causing clubbing of digits due to lowered oxygen levels in the blood); dehydration; liver disease; nasal polyps; sinusitis; infertility (especially male); acute pancreatitis; malabsorption; electrolyte disturbance.

#### *3.3.3.2.2 Targeted testing*

- The S*weat Test* has for many years been the standard diagnostic test to identify CF. Electrolyte levels of >60 mmol/l indicate a high probability of CF [6].
- *Transepithelial nasal potential difference* (TNPD) testing, like the sweat test, measures abnormal electrolyte levels, in this case in the respiratory epithelia.
- *Fecal elastase test* to detect steatorrhea.
- *Genetic screening* for CFTR mutations can identify up to 98.7% of sequence variants in the case of full sequencing [161] (which is still expensive). The more common targeted mutation analysis uses a panel of common mutations [162], the detection rate of which varies according to ethnic background [8].
- *Immunoreactive trypsinogen* testing as part of the Guthrie Test is commonly used in the US and Europe for neonatal screening [6].

#### *3.3.3.2.3 Inheritance*

- Family history
- Ethnicity – different CFTR mutations and mutation frequencies are found in various ethnic groups. This study used South African census data for 2018 as a starting point for the relative ratios between the various major ethnic groups [163].

#### *3.3.3.2.4 Other disease diagnoses*

- CF affects the pancreas; it is therefore not surprising that diagnoses of CF and diabetes often overlap.
- Especially in southern Africa, recurrent respiratory tract infections (one of the more common CF symptoms) regularly occur in many HIV- and tuberculosis-positive individuals; this represents a far larger percentage of the population than the relatively rare incidence of CF.

#### *3.3.3.2.5 Unobserved (latent) variables*

Latent parameters are to be identified – these are unobserved factors (also sometimes called "confounding variables") which may significantly affect the known parameters in the model, and thereby interfere with causal interpretation of the model structure. It is not necessarily expected that such latent parameters will be explicitly identifiable (mapping to measurable or known parameters); they may be abstract.

### 3.3.3.3 Accruing evidence

Of all the input data available, the single most accurate test is probably full genetic sequencing for CFTR mutations, with a sensitivity as high as 98.7% (at least in some population groups) [161]. For lack of data to the contrary, the specificity of this test is assumed to be the same value – this is unlikely to be accurate, but it can easily be updated when more credible information is obtained.

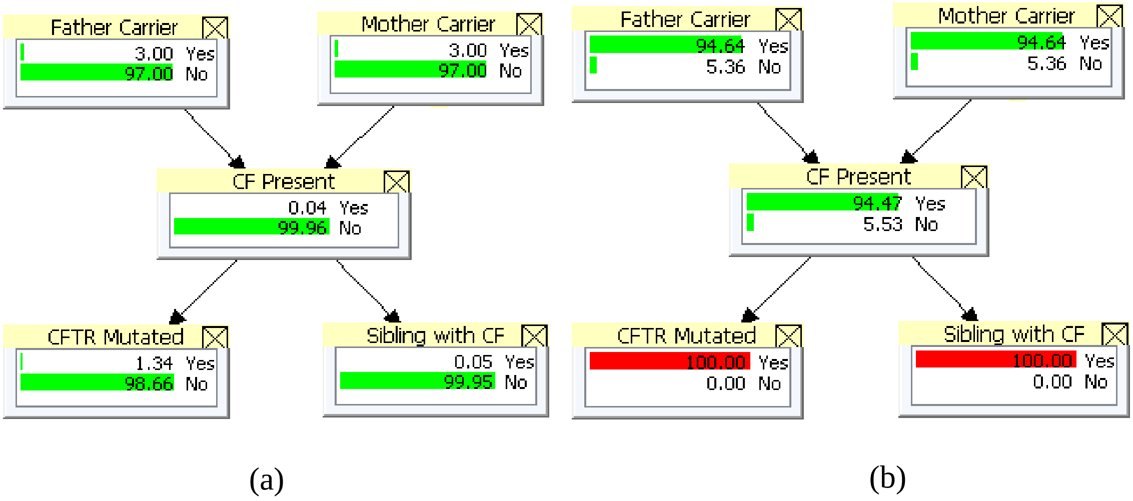


**Figure 3.1** – Simplified graphical model showing *a priori* degrees of belief (a), as well as the updated values (b) when a positive CFTR mutation sequencing result is combined with the presence of a sibling with CF (evidence shown in red).

Figure 3.1(a) shows a simplified graphical model for CF, using only the probabilities that a Caucasian person's parents are carriers of one mutated CFTR gene (3% each – for motivation see section 4.3.3.2), CFTR sequencing test results, and sibling CF status.

The chosen parental carrier probabilities yield a 0.04% chance (1 in 2500) that the subject would have CF, via Mendelian inheritance. This value, combined with the CFTR mutation sequencing test specificity, results in a 1.34% probability that a random person would test positive for CFTR mutations in both alleles, as shown in Figure 3.1(a).

#### *3.3.3.3.1 Adding information*

The graphical network now allows evidence to be entered, replacing the initial statistical values with actual knowledge as it becomes available. The effect of this evidence is propagated through the network according to the joint probability distribution functions connecting the nodes. Evidence can be added to any node.

Owing to the relative rarity of phenotypically manifest CF in the general population, a positive CFTR mutation sequencing result (i.e. reporting the presence of two mutant CFTR alleles in the same individual) on its own still only implies a probability of 2.95% that the patient actually has the condition.

This is the reason behind clinical diagnosis guidelines to combine the results of multiple screening and diagnostic tests. If, for example, the CFTR sequencing test result is combined with the presence of a sibling with CF, the probability that the patient has the condition goes up to 94.47% as shown in Figure 3.1(b).

Note that this information is also propagated upwards to the nodes denoting the parents, resulting in a dramatically increased expected likelihood (or *belief)* that they would be carriers of a mutated CFTR gene.

#### *3.3.3.3.2 Extended BN for CF*

Figure 3.2 presents a network for CF, incorporating all the input data mentioned in section 3.3.3.2 above. Each node in the graphical network can have two or more states, and the interconnections consist of the joint probability distribution functions of the connected node states. These functions are based on statistics (specifically the sensitivity and specificity of the various diagnostic tests, as far as they are known), expert opinion, genetic (Mendelian) theory, and sociological data, resulting in a distribution of node states which initially reflects the general population.

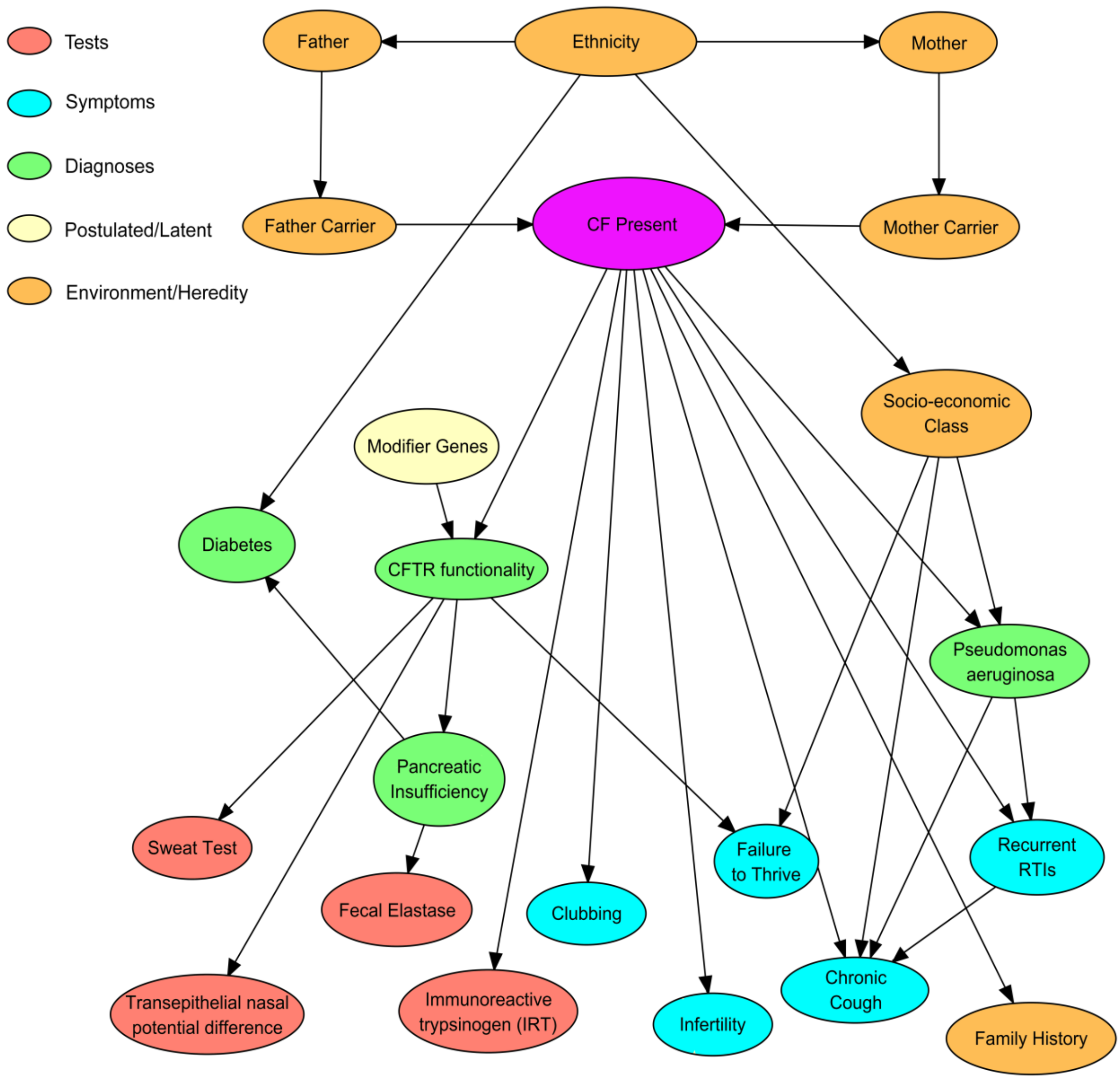


**Figure 3.2** – Graphical model of CF, incorporating observations (consisting of symptoms and test results), as well as family history, ethnicity, other disease diagnoses and latent variables.

The joint probability distribution functions in the graphical model of Figure 3.2 were populated with estimated data, which should be reasonably accurate for the South African context [163].

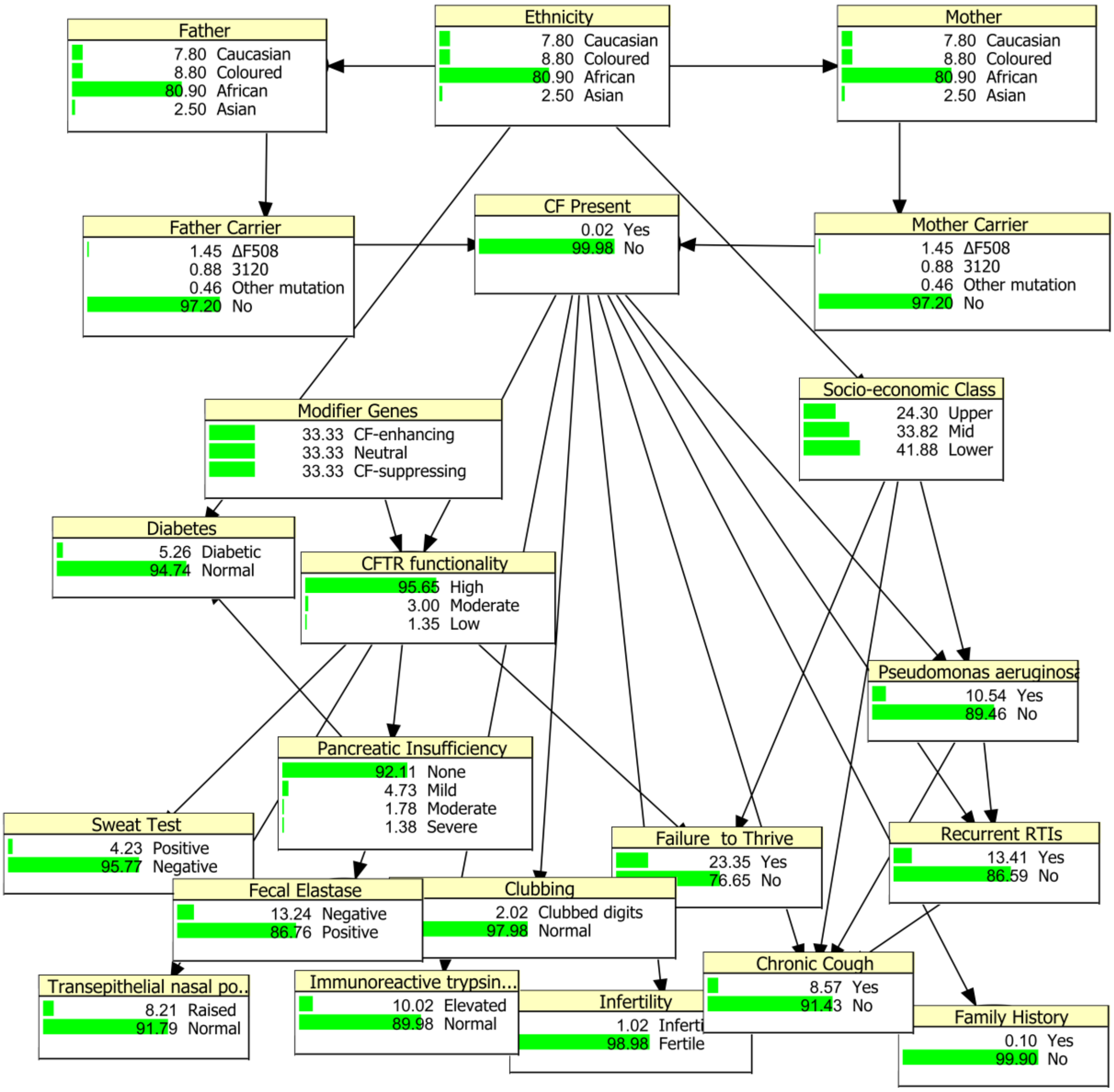


**Figure 3.3** – Default initialised state of CF model.

The incidence of the various CFTR mutations in the different ethnic groups is probably inaccurate, due to the dearth of data for especially non-Caucasian individuals [8]. In Figure 3.3 the green bars indicate the probabilities (expected occurrence rates) of the various states for each node. When a specific state becomes known, this can be set, collapsing the probabilities for that specific node to 100% of the known state, and 0% for the other(s).

Despite the questionable accuracy of the inputs, the model nonetheless plausibly predicts the incidence of CF in South Africa at approximately 0.02% of births; this jumps to more than 0.04% when ‘Ethnicity’ is set to Caucasian (i.e. European), once again agreeing with observed reality, where approximately 1 in 2500 Europeans are born with CF [33], [164].

The node labelled ‘Modifier Genes’ near the middle of the model acknowledges the probable presence of epistatic genes which can affect the functionality of the CFTR gene. For lack of information three equally likely possible states are defined (‘CF-enhancing’, ‘Neutral’, and ‘CF-suppressing’) – see Figure 3.3. In reality there are probably a large number of contributing factors, resulting in a continuum of possible modifying effects, ranging between the extremes as shown. This can be approximated by suitably adjusting these inputs as information becomes available.

In the middle, on the right-hand side of the model, another factor that is correlated with ethnicity (in South Africa) is shown. Socio-economic class in itself does not cause CF, but it is statistically different for the various ethnic groups, and several of the more important diagnostic observations (i.e. failure to thrive, and respiratory tract infections – RTIs) also tend to be less common among the higher socio-economic classes. The Bayesian model automatically takes this into account, if the inputs are known. This factor may become irrelevant once accurate genetic diagnosis is achieved in all ethnic groups.

As mentioned in section 2.2.2.2.2 above, the Bayesian combination of evidence as shown is only valid if the evidence is independent, i.e. when the tests measure unrelated factors. For example, the *Sweat test* and *Transepithelial nasal potential difference* (bottom left in Figure 3.2 and Figure 3.3) are both aimed at detecting atypical electrolyte levels, which are indicative of CFTR abnormality – when one of these tests comes back positive, another positive result on the other does not add significant additional evidence supporting an eventual CF diagnosis. However, confirmatory evidence would be important in the case of false positives and negatives.

### 3.3.4 Coding and data availability

The graphical models presented above were created in Hugin 8.2[13], and can be found at https://github.com/JohanViljoen/Graphical-Models.

## 3.4 CONCLUSION

### 3.4.1 Early detection

Even though the model presented here has limited diagnostic value due to the difficulty of obtaining accurate observational data, two important observations can be made regarding the use of a graphical model for diagnostic purposes:

- All tests and observations are imperfect, with non-zero false-positive and false-negative rates. Bayesian networks can accommodate such uncertainty rigorously.
- All diseases, including CF, present with a range of severity in different patients, which may vary from acute to imperceptible (in which case it will most likely not even be diagnosed).

Early detection and treatment of disease in general improves the outcome for most conditions, including CF [162]. Many First World countries therefore conduct neonatal testing for CF [6], [165], [166], [165].

### 3.4.2 Challenges

The following challenges remain:

- ***Determining a suitable graphical network structure***, reflecting the relationships and dependencies between nodes, as shown in the attempt presented in Figure 3.2.

[13] See https://www.hugin.com/

- ***Extracting reasonably accurate conditional dependence tables*** (the joint probability distribution functions connecting nodes) from CF studies and other databases. The problem here lies largely in the difficulty of obtaining validation data for healthy (at least non-CF) individuals, which is especially difficult in the case of the CF-specific and/or more expensive tests.
- ***Identifying latent variables*** which may affect and/or connect known parameters in the model, and potentially interfere with causal interpretation.

### 3.4.3 Discussion

The discovery of links and independences hidden in observed medical data may improve understanding and therefore also management of diseases, not only in individuals, but in populations. When suitably verified this could *inter alia* contribute to improved personal medical care and better public health management and policy, by

- facilitating causal predictions, such as the outcome of an intervention (treatment as well as the management of social and environmental factors),
- providing suggestions to patients and populations to manage elements of their environment which are causally linked to their diseases, symptoms and prognoses,
- estimating risk or uncertainty inherent in proposed interventions,
- weighing utility against risk,
- aiding the determination of policy aimed at improved resource allocation,
- and making truly personalised medicine feasible.

If the model can be suitably validated against observations, the generative property of Bayesian networks may enable causal prediction and simulated experimentation [150], which could complement medical interventions on human subjects.

# CHAPTER 4 POPULATION SIMULATION

## 4.1 INTRODUCTION

As alluded to in section 2.3, two important principles are pertinent when considering the establishment and dispersion of monogenic variations in a population: the selection coefficients conferred by the allele in homozygous and heterozygous forms respectively, as well as the size and structure of the population itself. Regarding the latter, much effort has been expended to analytically model the effects of local structure, migration, and inbreeding [9], [10], [12], [11]. As shall be demonstrated in the following sections, most of these 'correction factors' can be obviated, at least for human populations, by executing a stochastic model in which the breeding radius is a normal distribution [10] in social space (see Equations (2.3) and (2.4)), constrained to a probabilistic range motivated by the sociological work of Dunbar and Lehmann [13], [14].

## 4.2 DESIGN

### 4.2.1 Assumptions

While creating a numeric population simulation tool, a number of assumptions were made:

- A single autosomal genetic locus (with a mutation/variation that may affect the procreation probability of carriers) is considered. This reflects the situation in CF and other monogenic variations, including those shown in Table 2.2.
- Compared to the general (non-carrier) population, heterozygous and homozygous carriers of mutated genes can have different survival/fecundity rates. For example, homozygous CF results in a selective advantage $s_{hom}$ approaching -1.

- Individuals select mates for procreation purposes from a limited community $N_n$, which in general is much smaller than the size of the entire population. $N_n$ includes the effects of immigration and population structure, eliminating the need to estimate these factors.
- For a population of humans, Dunbar and Lehman motivate an upper cognitive limit to the number of stable social relationships that an individual can maintain [13], [14]. This is used to inform the realistic community size from which an individual can select a mate. Dunbar's number for humans is estimated to lie in the range of 100 to 230, with 148 being the nominal value. Values for $N_n$ of this size and smaller are considered to be reasonable for human populations.
- For modelling purposes, a two-dimensional grid lends itself well to processing and visualisation; this does not imply that human social networks (of the close, meaningful kind, as described by Dunbar, and specifically not the far more tenuous constructs found on social media) are necessarily two-dimensional in nature. However, the only relevant characteristic of the distribution is the distance $r$ as in equation (2.4). Irrespective of the dimensionality of the grid, the correlation between the Gaussian parental probability distribution functions of two individuals depends solely on the distance $r$ between them, scaled by the standard deviation $\sigma$. This correlation function, being the convolution of two normal distributions, is of course itself a normal distribution, with twice the variance of the parental probability function.
- Constant environmental conditions are assumed across the entire population – although provision can be made for different geographical conditions to reflect situations such as discussed above. All simulations presented in this document were conducted with this assumption of constant conditions in mind.
- In general the effective population number $N_e$ is assumed to be large and comparable to the total population census size $N$), and specifically much larger than $N_n$, i.e. $N_e >> N_n$. This is adjustable, however, to allow for exploration of effects in smaller populations and genetic isolates.

### 4.2.2 Structure

The following structure is used for the population itself:

- A two-dimensional array is created, with every element representing an individual in the population. Computational resources now make it feasible to create and process simulated populations consisting of millions, and even billions, of individuals.
- Each individual has one of four possible states: dead, no mutation (i.e. wild-type homozygous), heterozygous, or homozygous.
- The population array is closed upon itself, with edges wrapping around. This eliminates any edge effects that discontinuities may otherwise introduce.
- Physical proximity in the elements of the population array is used as a proxy for social closeness – i.e. an individual is more likely to breed with another nearby individual than with a remote one, according to a two-dimensional normal (Gaussian) probability distribution.
- The effective size of the community around each individual is changed by varying the standard deviation of the normal distribution, with $N_{\mathrm{n}}$ as in Equation (2.3). Also see section 4.3.1.4 and Figure 4.5 below.

### 4.2.3 Simulation procedure

For a given set of parameters, the following steps are executed:

1. Set population size $N$, community size $N_{\mathrm{n}}$, initial carrier prevalence, advantage/disadvantage for heterozygous and homozygous carriers ($s_{het}$ and $s_{hom}$), and *de novo* mutation probability.
2. Initialize the population randomly with a desired initial fraction of carriers.
3. For each individual in the population, change its status to dead with a probability dependent on its current status, to statistically reflect the selection coefficient associated with its status. This is done by computing a normalised survival probability, based on the ratios of the selection coefficients of the three relevant non-dead states (wild-type, heterozygous, or homozygous).

4. For each element of the population array:
    a. Randomly select two distinct non-dead parents from its community $N_n$, according to the proximity probability distribution as in Equation (2.4).
    b. Generate a status according to Mendelian inheritance probabilities from the two parents.
    c. Randomly introduce a *de novo* mutation with a specified (typically low and possibly even zero) probability.
5. Update population statistics and display.
6. Return to step 3.

### 4.2.4 Coding and data availability

The simulation was coded in Delphi. The software, instructions and raw data files are available on Github at https://github.com/JohanViljoen/PopSim.

## 4.3 RESULTS

### 4.3.1 Validation

#### 4.3.1.1 Establishment rate

To investigate the statistical fate of a mutation with selective advantage *s* (for both heterozygous and homozygous cases, as assumed by Haldane [114] and Barrett [121], the simulation tool was configured to introduce a single mutation in a virtual population (population size $N$ = 2.5x10$^5$), and then to cycle through the generations until the mutation either becomes ubiquitous or extinct.

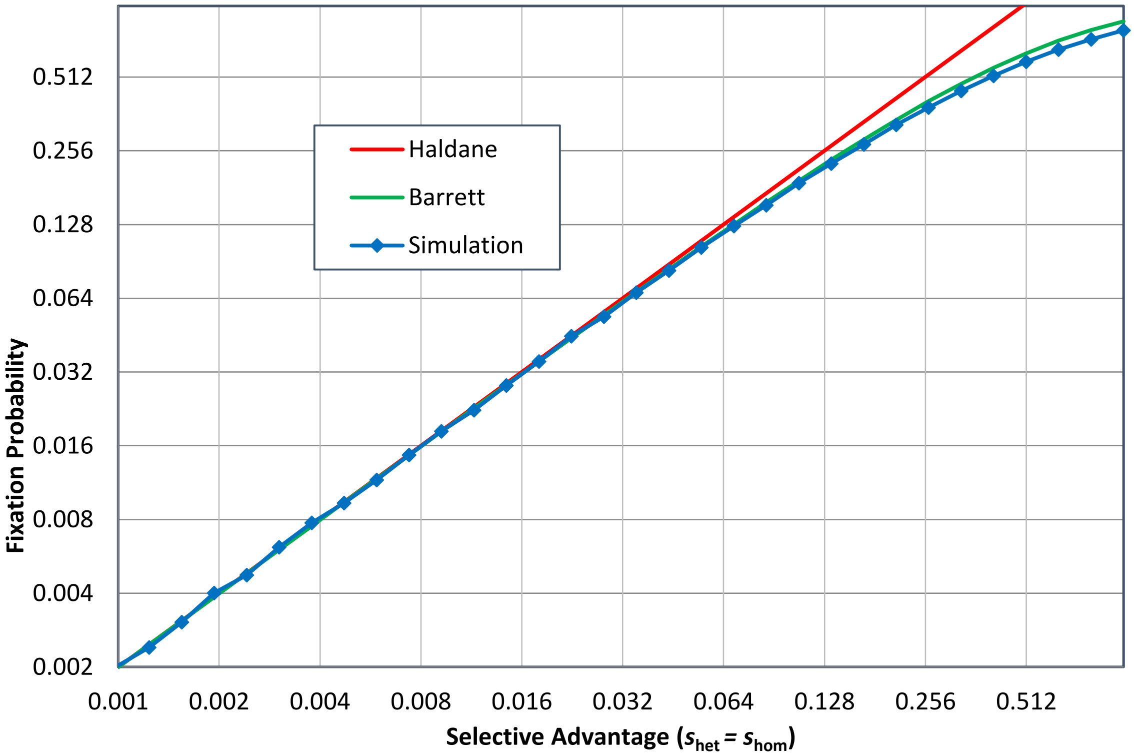


**Figure 4.1** – Establishment probability of a beneficial mutation. Fixation rate for a single mutation as a function of (positive) selective advantage in a panmictic population ($N$=$N_n$=2.5x10$^5$).

This was repeated 200 000 times for each point in the graph shown in Figure 4.1, for a total of ~112 million generations. The results confirm Haldane's prediction, including his *caveat* that it is only valid for small values of *s*. The stochastic simulation initially closely follows Haldane's $P = 2s$ line, but then gradually deviates from it as *s* increases, as predicted by Barrett's approximation in Equation (2.5). This is expected: the fixation probability cannot exceed 1; it can only approach unity asymptotically as *s* grows.

### 4.3.1.2 Equilibrium levels

In Addendum A a novel analytic derivation is presented, predicting the eventual equilibrium prevalence of a highly deleterious recessive variation such as CF (i.e. where $s_{hom} = -1$) as a function of the heterozygous selective advantage $s_{het}$.

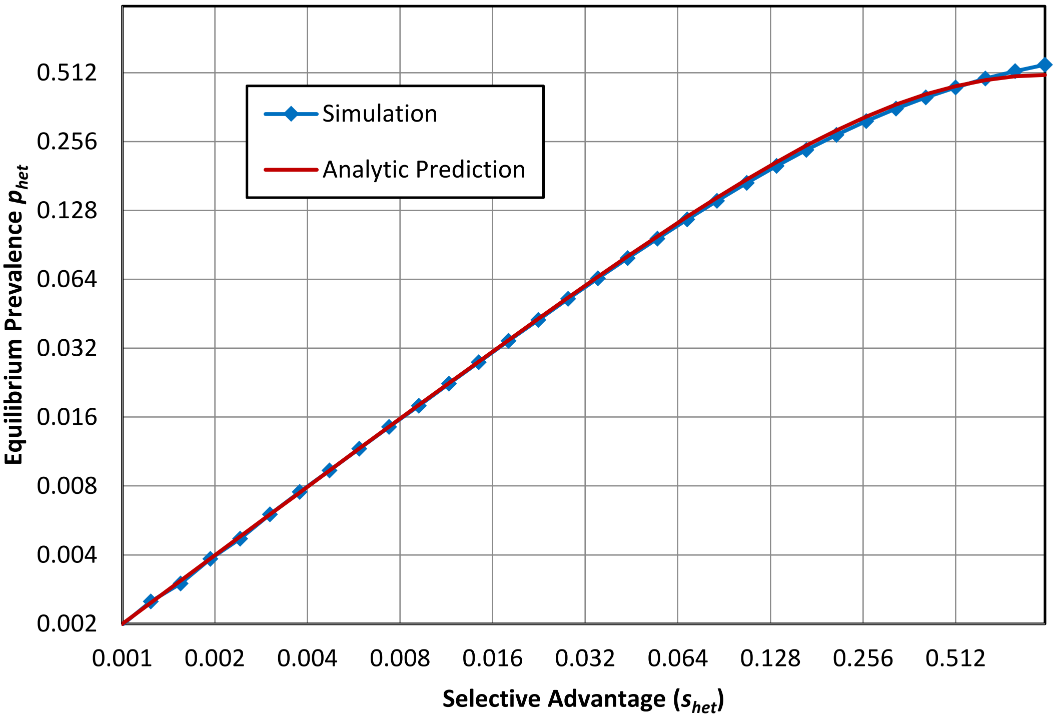


**Figure 4.2** – Equilibrium levels of a deleterious mutation. Equilibrium level as a function of (positive) heterozygous selective advantage, with homozygous selective advantage = -1 ($N$=4x$10^6$, $N_n$=$10^6$, $s_{hom}$=-1).

As with Figure 4.1, a stochastic simulation was executed, comparing the results with the predicted equilibrium levels as in Equation (A.1). Figure 4.2 demonstrates an excellent agreement even up to very high heterozygous selective advantages.

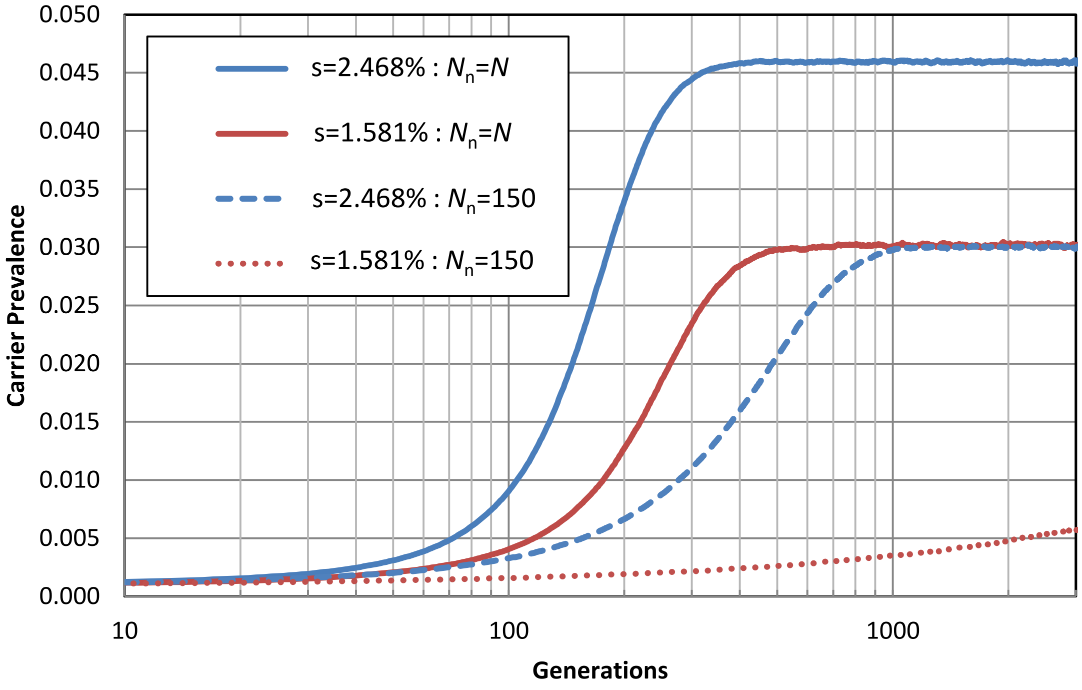


**Figure 4.3** – CFTR Carrier prevalence. Results for numerical simulations of mutation dissemination in a test case to stabilise at a CFTR mutation carrier frequency of 3%. $N$=$10^8$, with each curve the average of four independent simulation runs to 3000 generations.

Figure 4.3 was generated to test the numerical simulation, using two different selective advantage values *s* for heterozygous carriers, with homozygous carriers having a 100% disadvantage ($s_{hom}$=-1), i.e. none survive to procreate, which until quite recently approximated reality for CF.

In the Caucasian population the prevalence of heterozygous CFTR carriers is often presented as approximately 4% [33]. However, as will be demonstrated (see section 4.3.3.2 below), this is an overestimate: the actual value is probably closer to 3%, and may even be significantly lower, if equilibrium has not been reached, as explored in section 4.3.3.3. To stabilize at a prevalence of 3% requires either panmixia and $s_{het} = 1.581\%$, or, more realistically, a community size of 150 (approximately Dunbar's number for humans) and $s = 2.468\%$. Both of these scenarios eventually approach heterozygous carrier equilibrium at 3%, after hundreds of generations.

#### 4.3.1.3 Dispersion

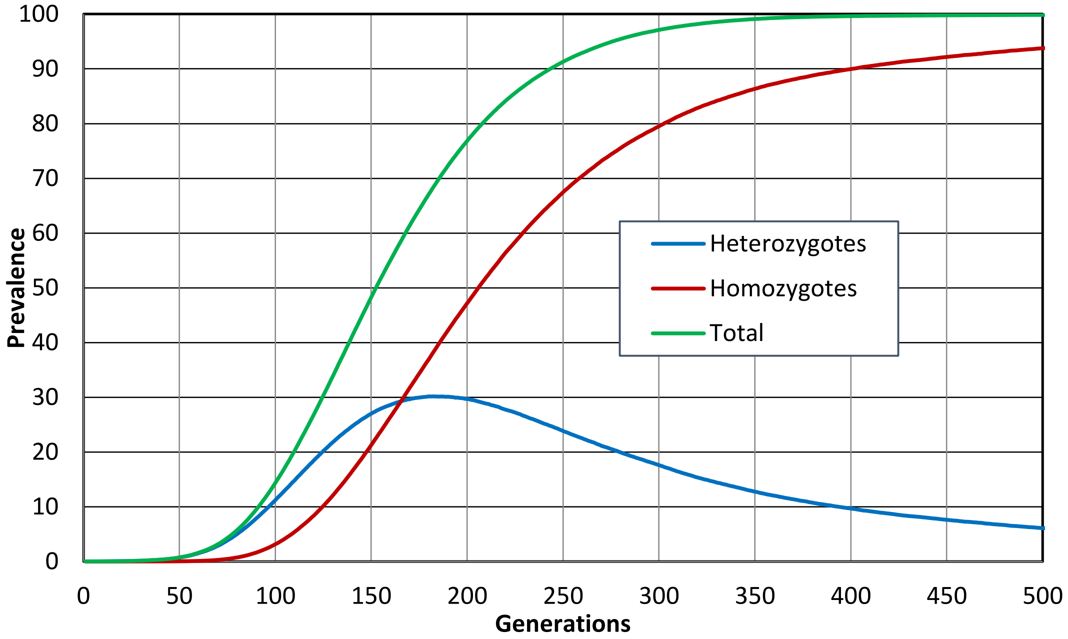


**Figure 4.4** – Lactase persistence. The spread of a purely beneficial mutation ($s_{het}$=$s_{hom}$=0.1, $N$=5x10$^6$, $N_n$=150).

Lactase persistence is an autosomal-dominant inherited genetic trait [167], [168], [169] associated with the LCT gene (MIM 603202) and is especially prevalent in European populations, with evidence of strong recent selection during the last 5000-1000 years [170], coinciding with the domestication of cattle and a rise in dairy farming. In such a setting, the ability of adults to derive nutrition from dairy confers an obvious advantage.

Bersaglieri *et al* estimate the selection coefficient for lactase persistence to be in a range from 0.09 to 0.19 for the Scandinavian population (where the prevalence of the -13910T allele currently exceeds 80%) [170]. In Figure 4.4 a conservative value of 0.1 (10%) was assumed as the selective advantage for both heterozygotes and homozygotes. Starting from an initial low base, the prevalence of heterozygotes grows to a maximum value of just over 30%, by which time it has already been surpassed by homozygotes, which asymptotically approach 100%, given enough time (generations). After just more than 200 generations, 80% of the population carries the mutation – at a nominal 29 years per generation [171] this corresponds to 5800 years, which is near the lower end of the antiquity estimate for dairy farming. However, it seems reasonable to assume that dairy farming also took a long time to become common, which implies that the availability of milk, and consequently the effective selective advantage of LCT, gradually increased to its current value.

#### 4.3.1.4 Effective community size

In Equation (2.3) Wright defines the effective size of a two-dimensional Gaussian distributed neighbourhood as a circle with radius $2\sigma$ [116].

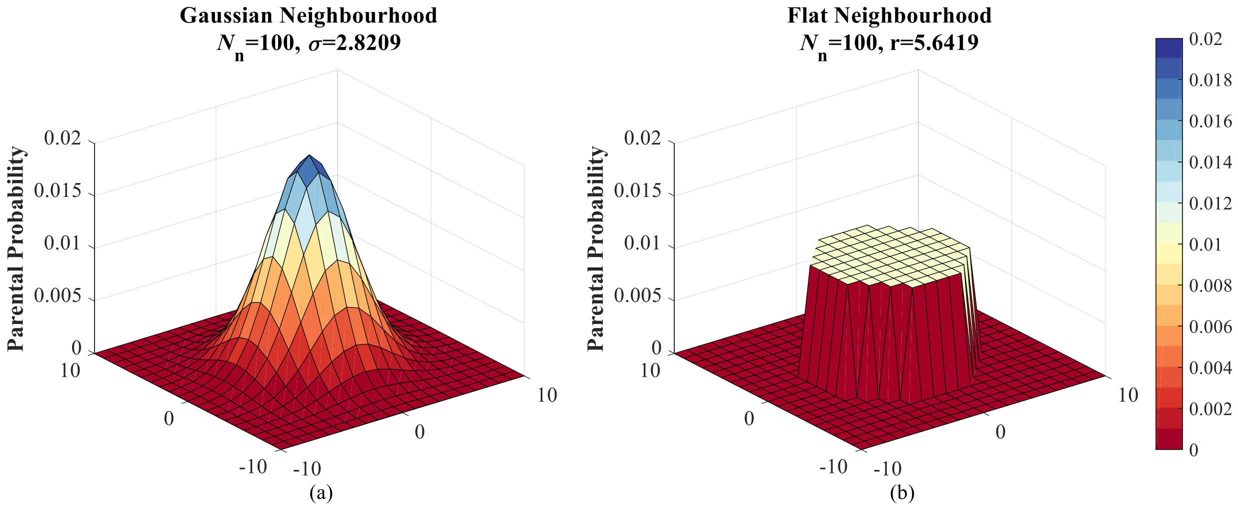


**Figure 4.5** – Gaussian and Flat neighbourhood definitions. (a) Gaussian distribution for $N_n$=100 → $\sigma$=2.8209, (b) Flat distribution: $N_n$ = the 100 nearest individuals.[14]

[14] The hue and intensity mapping used in this and subsequent plots has been designed to improve accessibility for individuals with colour-deficient vision [172], which is yet another ostensibly deleterious genetic variation that has nonetheless achieved significant levels in humans, implying some concomitant selective advantage.

To test this claim, the simulation was configured with a lethal recessive variation ($s_{hom}$=-1) for several different positive values of $s_{het}$, comparing the equilibrium carrier prevalence levels that are reached in a large population for a range of neighbourhood sizes $N_n$ when using either a Gaussian distribution as defined (see Figure 4.5(a)), or simply a flat distribution as shown in Figure 4.5(b), in which each of the $N_n$ closest neighbours are equally likely to be selected as parents.

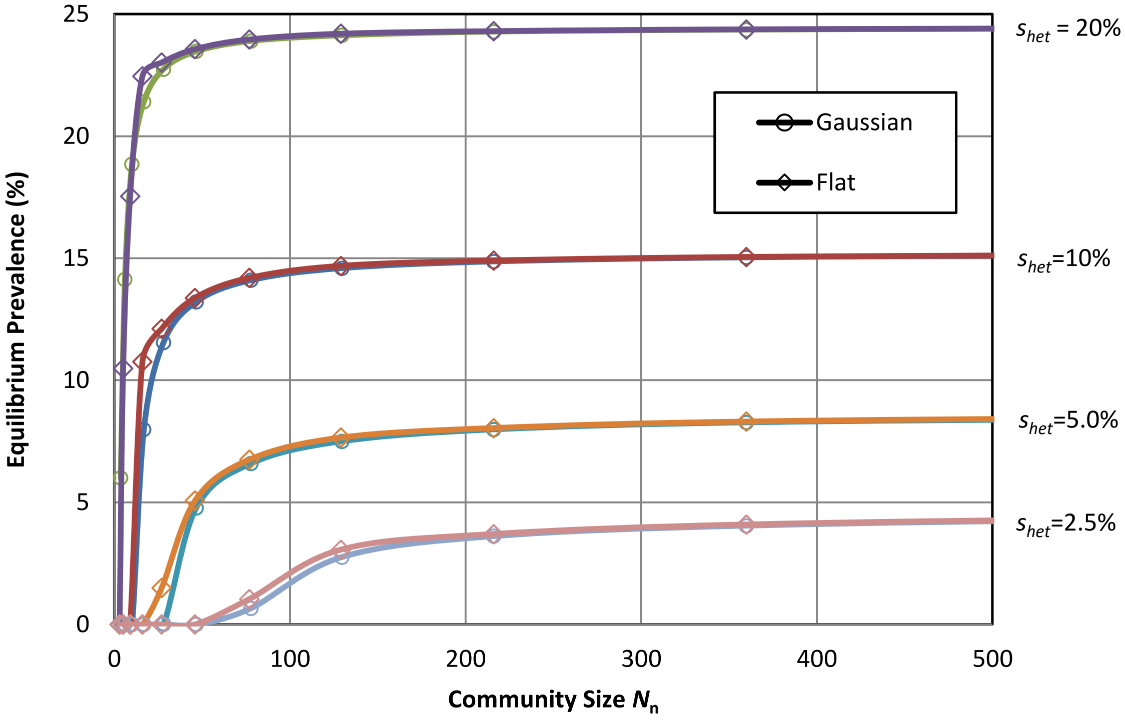
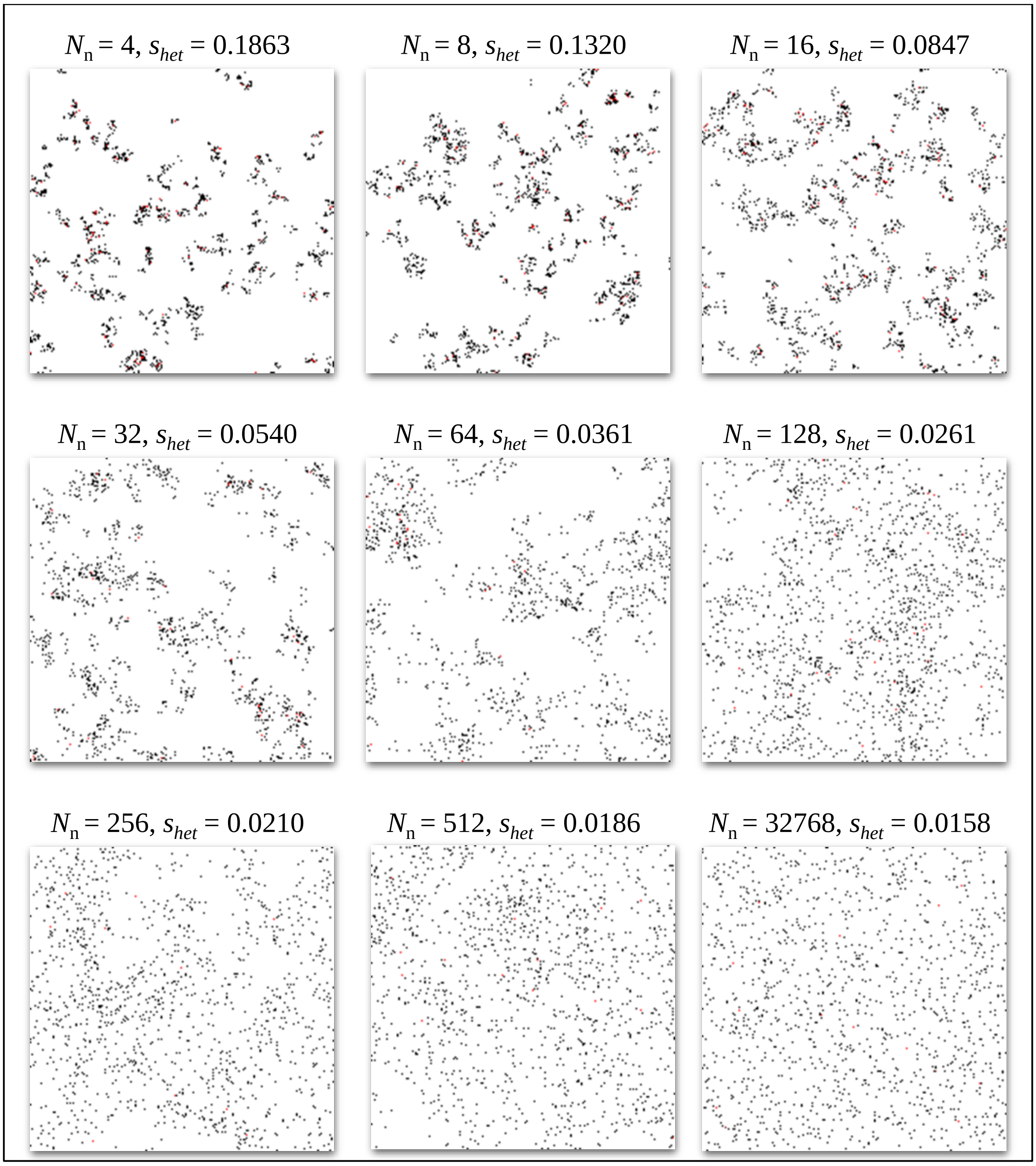


**Figure 4.6** – Comparison of equilibrium levels for Gaussian and flat neighbourhoods for various heterozygous selective advantages. ($N$=2.5x$10^7$, $s_{hom}$=-1).

As shown in Figure 4.6, the results match quite closely, except for very small values of $N_n$. This is not unexpected, because a flat distribution requires quantisation – only an integer number of individuals can constitute a neighbourhood where all are equally likely to be selected, and the resultant rounding errors therefore become significant when $N_n$ is small. In the interests of improved accuracy, a Gaussian neighbourhood was therefore used in all other simulations presented in this document.

### 4.3.2 Population structure

The neighbourhood shape (shown in Figure 4.5) determines the likelihood that one individual will breed with another at a given distance.



$N_\mathrm{n} = 4$, $s_{het} = 0.1863$

$N_\mathrm{n} = 8$, $s_{het} = 0.1320$

$N_\mathrm{n} = 16$, $s_{het} = 0.0847$

$N_\mathrm{n} = 32$, $s_{het} = 0.0540$

$N_\mathrm{n} = 64$, $s_{het} = 0.0361$

$N_\mathrm{n} = 128$, $s_{het} = 0.0261$

$N_\mathrm{n} = 256$, $s_{het} = 0.0210$

$N_\mathrm{n} = 512$, $s_{het} = 0.0186$

$N_\mathrm{n} = 32768$, $s_{het} = 0.0158$

**Figure 4.7** – Spatial structure of population at 3% heterozygous equilibrium prevalence level. Black = heterozygous, Red = homozygous, $s_{hom}$=-1.

When the effective $N_\mathrm{n}$ is small, this probability drops off very sharply with distance; mating with proximate individuals is far more likely than with remote ones. This leads, unsurprisingly, to local structure in the distribution of alleles.

When considering human populations, the space in question remains the social plane as introduced in section 2.3.1. Examples of representative emergent spatial structures at various community sizes $N_\mathrm{n}$ are shown in Figure 4.7, where the attendant $s_{het}$ was adjusted each time to result in the same heterozygous equilibrium prevalence level of 3%, for a recessive variation that is highly deleterious in the homozygous state ($s_{hom}$=-1), like CF. As $N_\mathrm{n}$ increases, the spatial distribution of alleles becomes progressively more homogenous, with the final panel closely approximating the panmictic situation.

### 4.3.3 Homozygous prevalence

#### 4.3.3.1 Constant heterozygous prevalence

To maintain a specific prevalence level of a CF-type variation in a population requires an increasing $s_{het}$ for decreasing values of $N_\mathrm{n}$, as seen in Figure 4.3 and Figure 4.7.

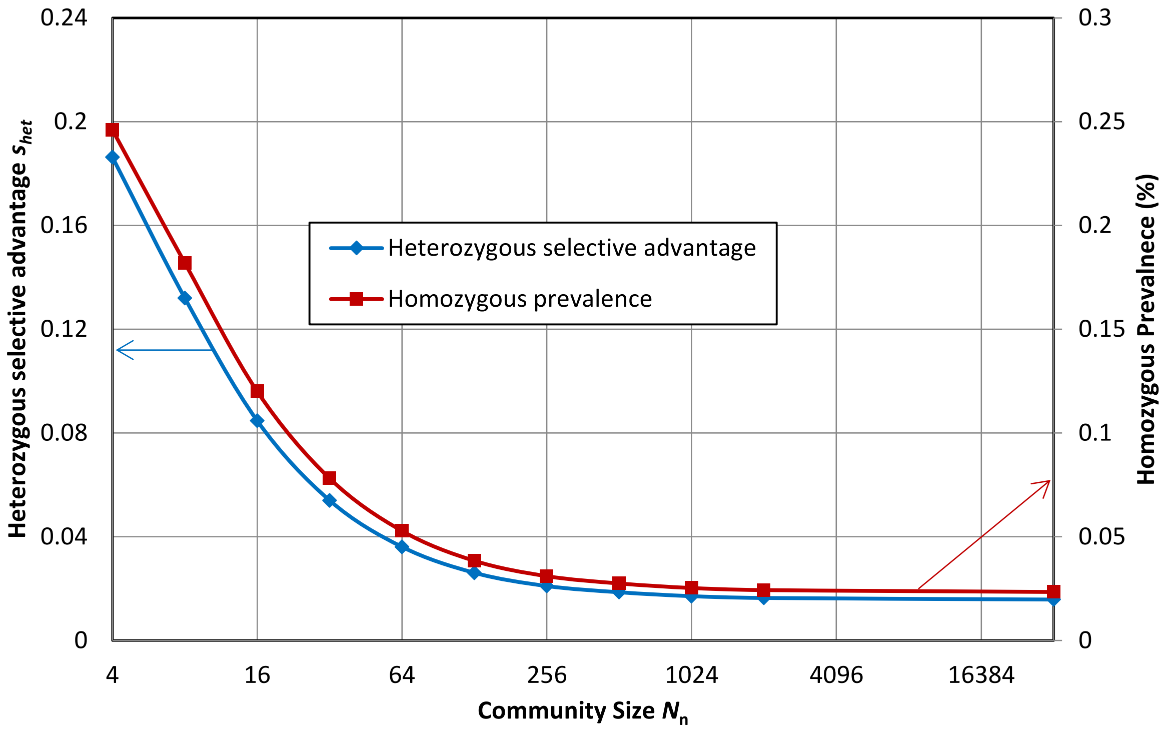


**Figure 4.8** – Heterozygous selective advantage $s_{het}$ and homozygous prevalence $p_{hom}$ *vs* community size $N_n$ at 3% heterozygous equilibrium prevalence level. ($s_{hom}$=-1.)

It also appears as if the prevalence of homozygosity follows a similar trend, which is confirmed when the actual values are displayed on the same graph – see Figure 4.8.

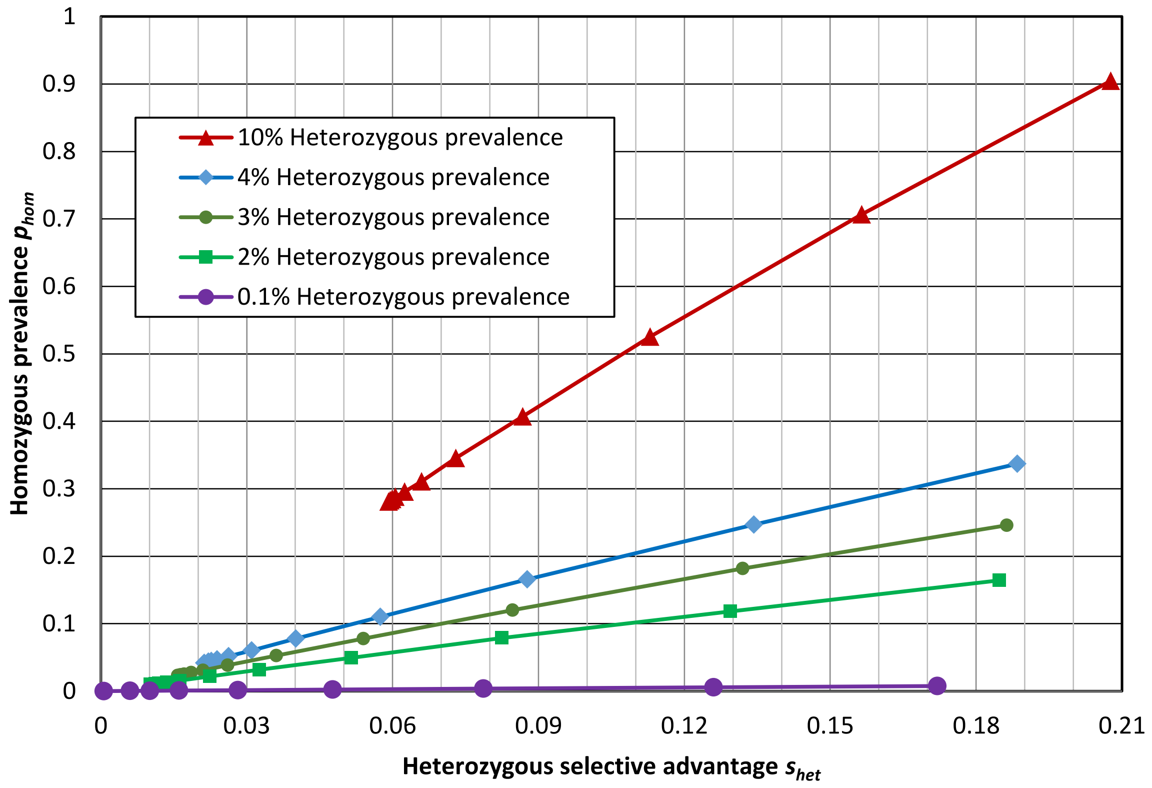


**Figure 4.9** – Homozygous prevalence vs. heterozygous selective advantage for various equilibrium prevalence levels. ($s_{hom}$=-1.)

When the homozygous prevalence is plotted as a function of $s_{het}$ for various heterozygous equilibrium prevalence values (including the 3% values shown in Figure 4.8) as in Figure 4.9, two observations can be made:

- Approximately straight lines are produced for each heterozygous equilibrium value.
- The homozygous prevalence seems to scale with the heterozygous equilibrium value.

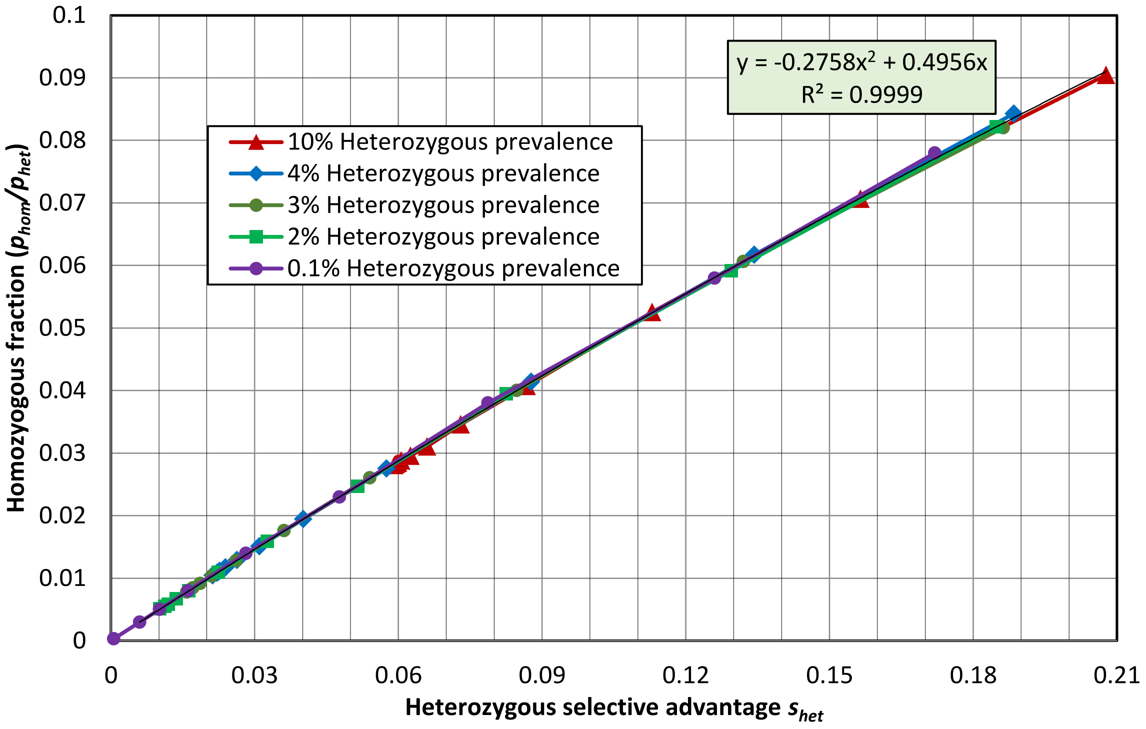


**Figure 4.10** – Homozygous fraction $p_{hom}/p_{het}$ *vs* heterozygous selective advantage $s_{het}$. ($s_{hom}$=-1.)

Normalising the homozygous prevalence by dividing it by the heterozygous equilibrium prevalence value as suggested by the second observation above yields the graph shown in Figure 4.10. A second-order polynomial fitted through all the points (including the origin) results in the following empirical expression (R$^2$=0.9999):

$$\frac{p_{hom}}{p_{het}} \approx -0.2758 s_{het}^2 + 0.4956 s_{het}. \tag{4.1}$$

For small values of $s_{het}$ Equation (4.1) reduces to

$$p_{hom} \approx \frac{p_{het} s_{het}}{2}, \tag{4.2}$$

or

$$s_{het} \approx \frac{2 p_{hom}}{p_{het}}. \tag{4.3}$$

#### 4.3.3.2 Practical implications

The Republic of Ireland has the highest reported incidence of CF, at 1 per 1353 live births with an estimated 1 in 19 (5.26%) Irish people being heterozygous carriers [165]. For the USA the corresponding numbers are 1 CF case per 3200 births, with approximately 1 in 29 people (3.45%) expected to be heterozygous carriers. Substituting these numbers into Equation (4.3) results in the Irish $s_{het} = 0.028$ and the USA $s_{het} = 0.018$. It is noted, however, that the claimed carrier prevalence numbers are suspiciously close to the results that would be obtained by merely computing it from simple Mendelian inheritance assumptions ($p_{het} = 2\sqrt{p_{hom}}$).

When a stochastic simulation is executed with community size $N_n = 150$, it is found that, to reach an equilibrium at the Irish $p_{hom}$ of 1 in 1353, $s_{het} = 0.0329$, under which conditions the heterozygous prevalence $p_{het} = 4.6\%$, instead of the estimated 5.26%.

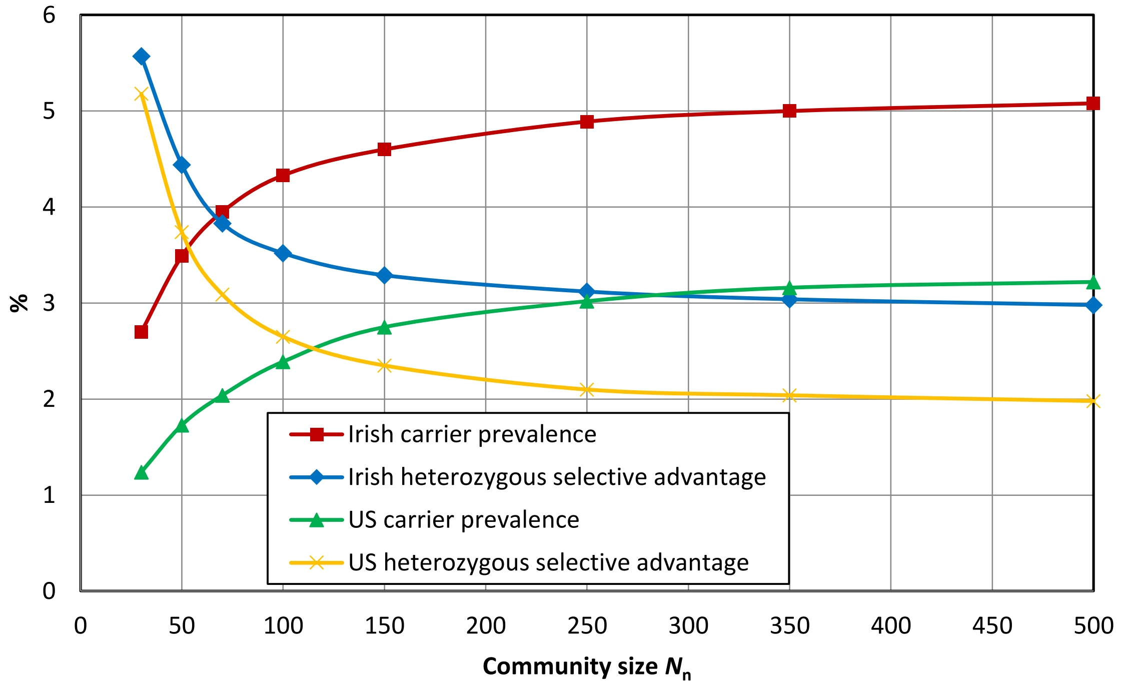


**Figure 4.11** – CF in Ireland and the USA. Heterozygous selective advantage $s_{het}$ and carrier prevalence $p_{het}$ as a function of community size for a CF incidence $p_{hom}$ of 1 in 1353 (Ireland), and 1 in 3200 (USA).

**Table 4.1** – Updated CFTR carrier prevalence level estimates for selected European and American populations, at effective community sizes of 150 and 100.

| **Population** | $p_{hom}$ | $p_{het}$ **panmictic** | $p_{het}$ **(*N*$_n$=150)** | $p_{het}$ **(*N*$_n$=100)** | **Reduction** |
|---|---|---|---|---|---|
| Ireland | 1:1353 | 5.44% | 4.60% | 4.33% | 14% –20% |
| UK | 1:2415 | 4.07% | 3.27% | 2.93% | 19% – 26% |
| USA | 1:3200 | 3.54% | 2.75% | 2.39% | 22% – 31% |
| Sweden | 1:5600 | 2.67% | 1.87% | 1.59% | 30% – 40% |
| Hispanic | 1:9200 | 2.09% | 1.28% | 1.05% | 39% – 50% |
| African-American | 1:15100 | 1.63% | 0.86% | 0.69% | 47% – 58% |
| Asian-American | 1:35100 | 1.07% | 0.40% | 0.32% | 62% – 70% |

When this situation is explored as shown in Figure 4.11, it is found that the actual heterozygous CFTR carrier prevalence is almost certainly lower than claimed, which in turn implies an increase in the heterozygous selective advantage. Even at the upper community size of 150, the values for $p_{het}$ are 4.6% and 2.75% for Ireland and the USA respectively, dropping to 4.33% (Ireland) and 2.39% (USA) at a community size of 100.

This suggests that the actual heterozygous CF carrier level in the Caucasian population is at least 20% lower than hitherto generally expected. Furthermore, the lower the actual incidence, the larger the discrepancy between the commonly used panmictic prevalence calculation and the results shown in Table 4.1. This has far-reaching implications for the carrier prevalence levels in population groups where the condition is less common, as well as for other relatively rare recessive deleterious variations such as those listed in Table 2.2.

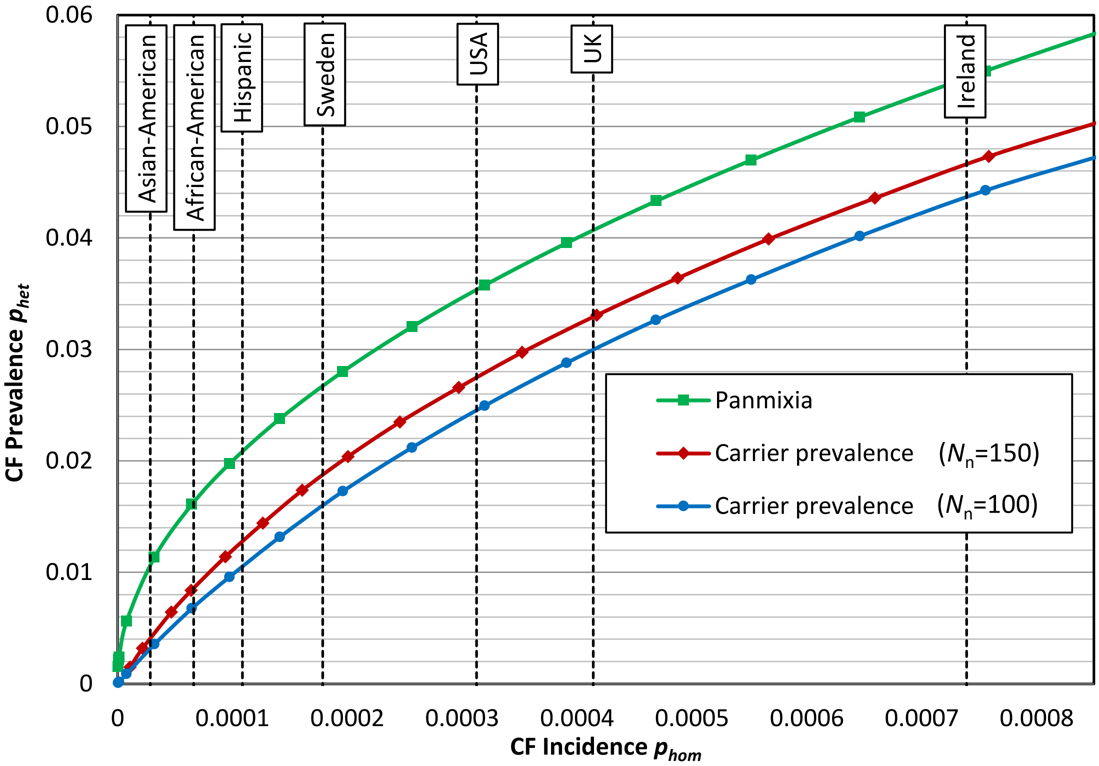


**Figure 4.12** – CF prevalence vs incidence for panmictic conditions and for community sizes of $N_n$=150 and $N_n$=100. Average of multiple equilibrium runs per point ($N$=2.5x$10^7$, $s_{hom}$ = -1).

Figure 4.12 presents a graphical depiction of the data in Table 4.1, illustrating the reduction in prevalence relative to the panmictic calculation.

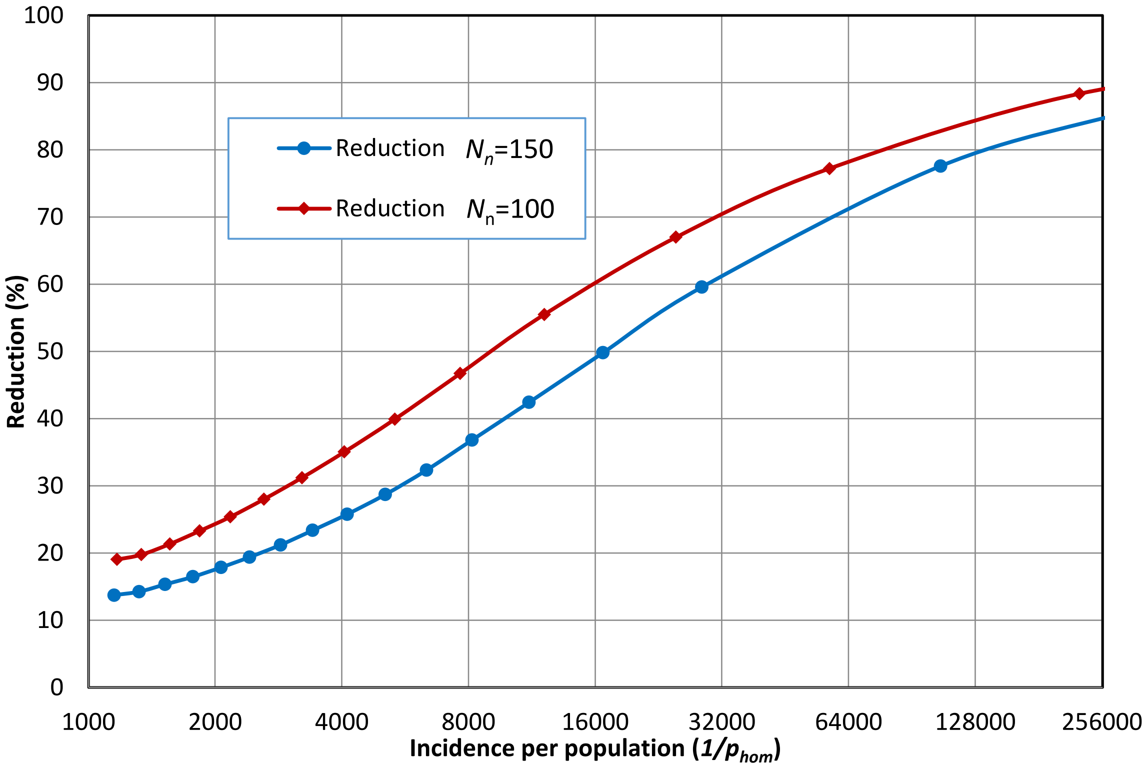


**Figure 4.13** – Prevalence reduction relative to panmictic conditions for $N_n$=150 and $N_n$=100. Average of multiple equilibrium runs per point ($N$=2.5x$10^7$, $s_{hom}$ = -1).

When the relative prevalence reduction values are plotted against the per-population incidence as in Figure 4.13, it becomes clear just how dramatically the carrier prevalence is overestimated when using a panmictic assumption, especially when the homozygous disease incidence is low, as is the case for many deleterious recessive diseases.

The preceding analysis is only valid for the case where the homozygous selection coefficient $s_{hom}$→-1, as was the case for CF until recently. Many of the conditions listed in Table 2.2 have homozygous selection coefficients somewhat larger than this, implying that a non-zero fraction of homozygous individuals may survive and be capable of procreation. If a realistic estimate of the applicable homozygous selection coefficient can be obtained, the same process that was followed above may be used to determine a credible carrier prevalence.

#### 4.3.3.3 Non-equilibrium

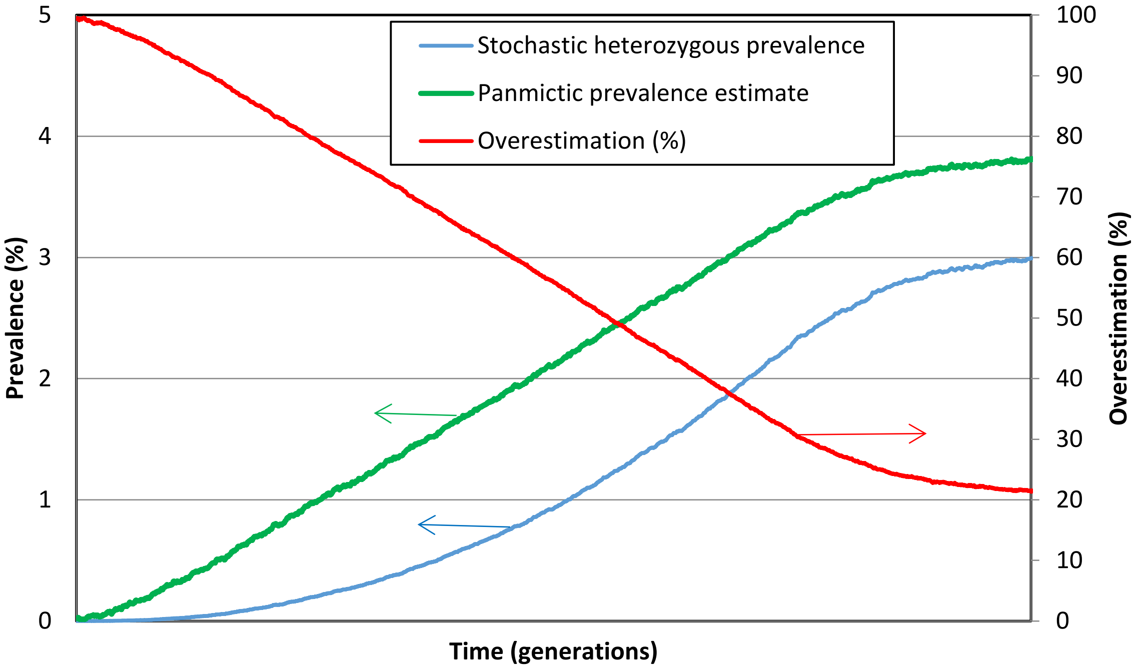


**Figure 4.14** – Heterozygous prevalence during mutation establishment. Comparison of stochastic heterozygous prevalence with panmictic calculation ($N$=1x10$^8$, $N_n$=150, $s_{het}$=0.02468, $s_{hom}$=-1).

When a mutation such as CF spreads through a large population, with selection coefficients as shown in Figure 4.14, it initially expands outwards from the site of introduction at a relatively low rate. As the effective circumference of the affected part of the population increases, the growth rate also accelerates. Eventually saturation levels are approached, at which point the rate diminishes until equilibrium is reached. This same behaviour can be observed in Figure 4.3 and Figure 4.4.

While such an allele is still rare, the incidence of homozygous individuals is of course very low. If, as is the normal practice, this homozygous incidence is used to compute an estimate of the heterozygous prevalence using the panmictic assumption, the results are seen to initially greatly overestimate the actual heterozygous prevalence resulting from the stochastic simulation, with the error eventually reducing to the values shown in Table 4.1, Figure 4.12, and Figure 4.13.

In other words, when equilibrium has not been reached, a computed estimate of heterozygous carriers is likely to be even more inaccurate than shown in section 4.3.3.2 above.

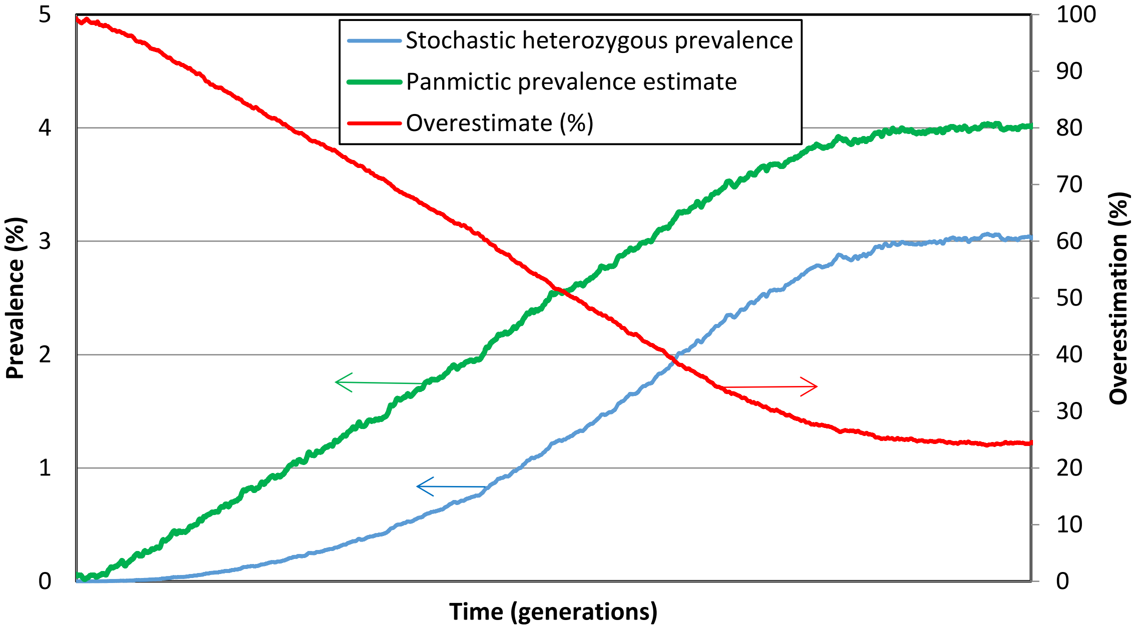


**Figure 4.15** – Establishment of a less deleterious variation. Comparison of stochastic heterozygous prevalence with panmictic calculation ($N$=3.6x$10^7$, $N_n$=150, $s_{het}$=0.0135, $s_{hom}$=-0.5).

Not all of the conditions listed in Table 2.2 are quite as lethal in homozygous form as CF was until recently. Arbitrarily assuming $s_{hom}$ = -0.5, and repeating the establishment experiment as for Figure 4.14 results in $s_{het}$ = 0.0135, to ensure an eventual heterozygous equilibrium of ~3%. Significantly, the behaviour as seen in Figure 4.15 is quite similar to that of the completely lethal variation considered earlier, leading to the conclusion that the carrier prevalence levels of less-lethal but still deleterious recessive mutations are probably overestimated to a similar degree as for CF.

In Europe the *F508del* variant of the CFTR gene is by far the most common allele associated with CF. According to the European Cystic Fibrosis Society (ECFS), this variant was found in 82.4% of 43190 European CF patients seen in 2016 [166]. Half of these (41%) were homozygous for *F508del*.

To determine the prevalence of CFTR mutations in the general population, a list of CF-associated variations of the CFTR gene (according to the CFTR2 mutations database at http://www.umd.be/CFTR/W_CFTR/gene.html) was compiled from the ExAC database [173] and the even larger gnomAD [174]. More than half of the humans in these databases are classified as '*European Non-Finnish*', which should correspond reasonably closely with the demographics covered by the ECFS. It is found that *F508del* is indeed the most common variation, occurring in 1.06% (ExAC) or 1.24% (gnomAD) of Europeans. This represents less than half of the individuals found to carry a CFTR mutation.

If the European population were panmictic, it would be expected that approximately 4% of European individuals would be carriers of CFTR variants, instead of the 1.92% found in the ExAC database, or 1.84% in gnomAD. For Hispanic (Latino) individuals the corresponding numbers are 2.06% (expected) versus 0.89% (ExAC) or 1.07% (gnomAD). These numbers are slightly lower than the predicted carrier prevalence values for $N_n$=100, as shown in Table 4.1.

These observations support our results by demonstrating a significant overestimation of CF carriers, while at the same time indicating that the effective $N_n$ for humans seems to be approximately 100, or even slightly less. It also expected that the ever-increasing amounts of genomic data currently being gathered will similarly reveal fewer carriers than currently expected in other recessive conditions, including those listed in Table 2.2.

### 4.3.4 Monte Carlo analyses

The real power of the stochastic simulation tool lies in its Monte Carlo functionality, which facilitates the automatic execution of multiple runs, while varying the starting conditions and input parameters. This enables the compilation of statistical results over millions of trials.

#### 4.3.4.1 Establishment of deleterious variations

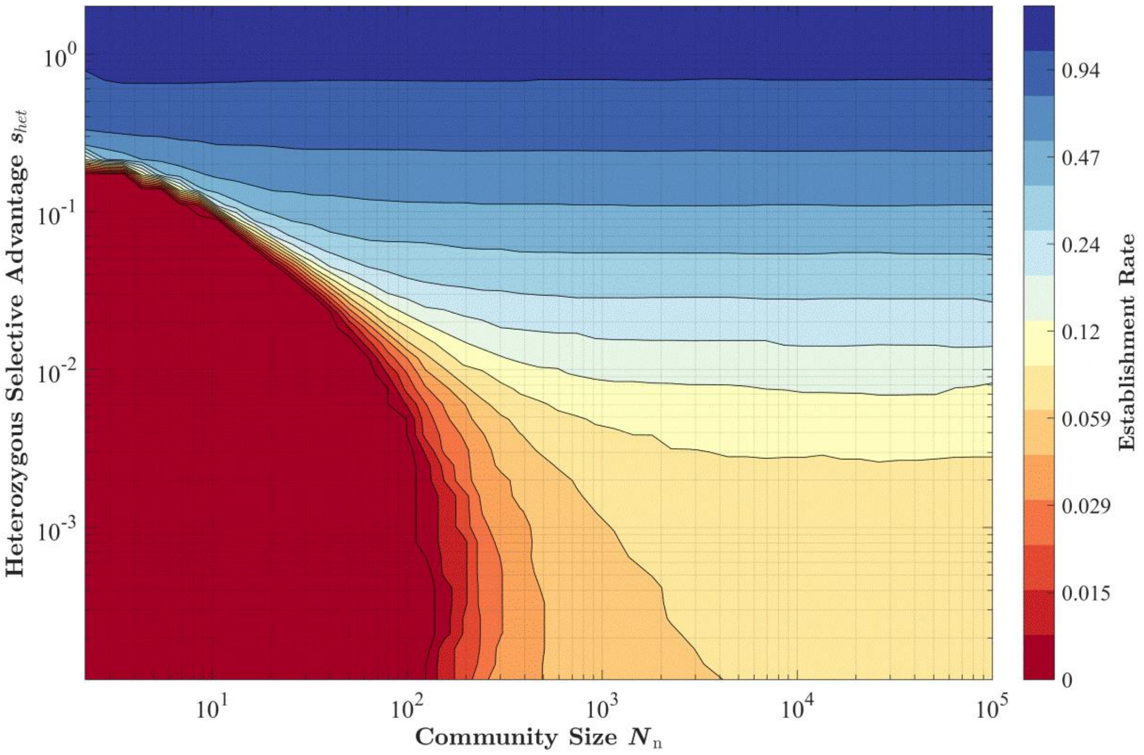


**Figure 4.16** – Rate of establishment as a function of community size and heterozygous selective advantage. Monte Carlo run results ($N$=$10^6$, $s_{hom}$ = -1).

Figure 4.16 shows the probability that a single mutation will indeed become established, as a function of community size and the heterozygous selective advantage that it confers, while keeping the homozygous selection coefficient equal to -1, which approximates the case found in diseases such as CF. From the data generated by the simulations underlying this plot one can also extract statistics regarding the average survival (in generations until extinction, if this happened) and maximum prevalence that a given mutation attained.

#### 4.3.4.2 Equilibrium levels

Figure 4.17 illustrates how the eventual equilibrium prevalence of a mutation (with homozygous selection coefficient $s_{hom}$ = -1) depends on both the community size and the heterozygous advantage that it confers on a carrier.

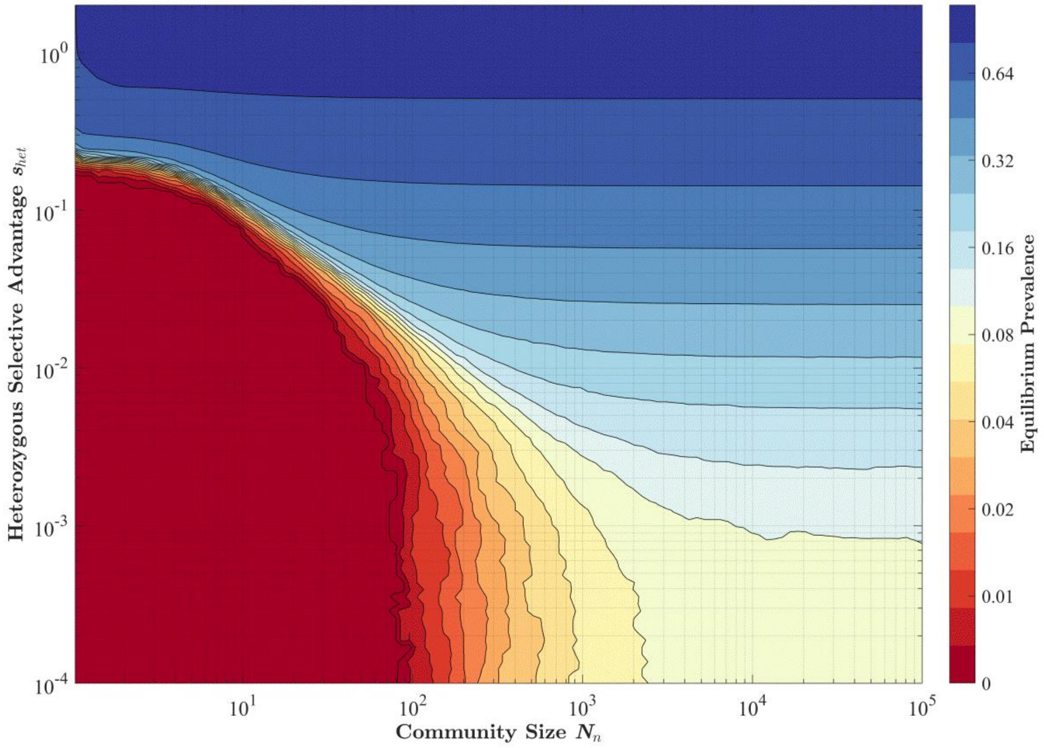


**Figure 4.17** – Equilibrium prevalence as a function of community size and heterozygous selective advantage. Monte Carlo run results ($N$=$10^6$, $s_{hom}$ = -1, max 10000 generations per point).

Figure 4.2, presented earlier, is valid for large values of $N_n$ and therefore corresponds to a vertical cross-section of Figure 4.17 on the right-hand side, while Figure 4.6 consists of four horizontal cross-sections through Figure 4.17, at the indicated heterozygous selective advantage values.

### 4.3.4.3 Underdominance

The preceding analyses mostly focused on overdominant mutations, i.e. where the heterozygous genotype enjoys a selective advantage over both the homozygous states (mutated and wild-type). Underdominance, also known as 'negative overdominance' [175], [176] is the situation where a heterozygous disadvantage exists, compared to either homozygous state.

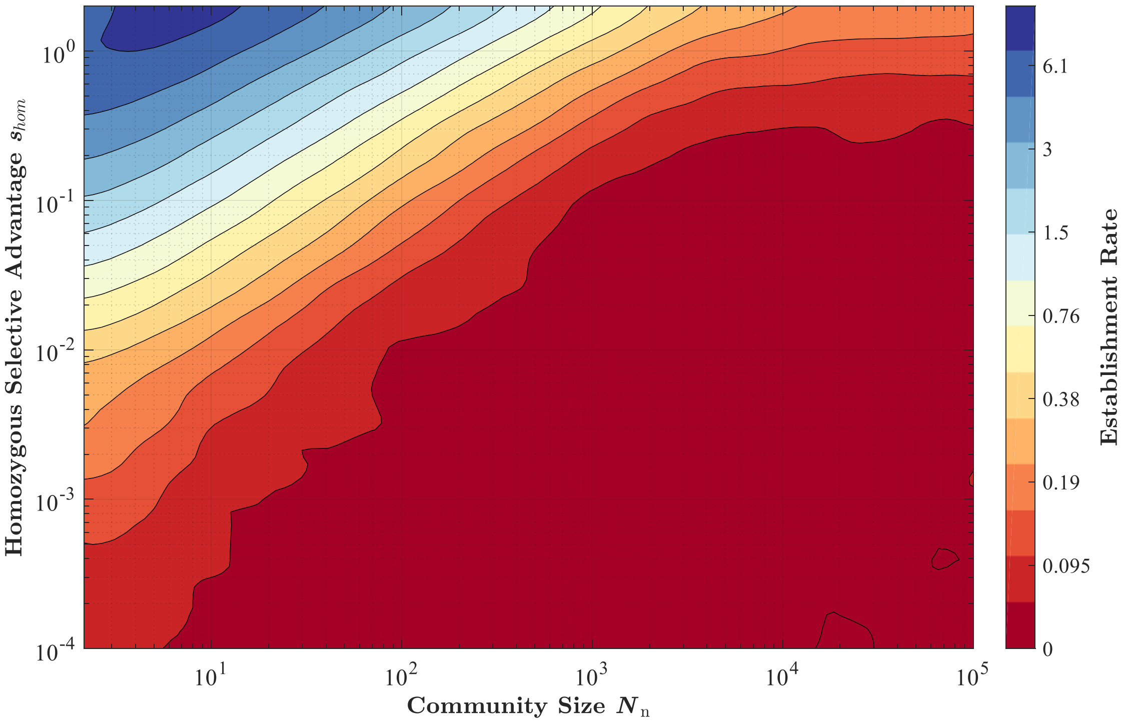


**Figure 4.18** – Establishment rate of an underdominant mutation as a function of community size and homozygous selective advantage. Monte Carlo run results ($N$=2.5x$10^5$, $s_{het}$ = $s_{wild}$ = 0, total number of generations : 3.377x$10^9$).

In humans the Rh factor, which is associated with an increased risk of haemolytic disease in the newborn, is subject to such selective pressures [177]. The mechanism is also used commercially to artificially introduce refractory genes into pest populations [176], [178], although in this case the establishment probability is not left to chance, as would be the case when a random mutation with potentially underdominant characteristics were to occur.

When a selective advantage only manifests in the mutated homozygous case, a single mutation is unlikely to become established, due to the effects of genetic drift, which is apt to extinguish it before it becomes sufficiently widespread to result in natural selection causing the homozygotes to outcompete the wild type. Figure 4.18 illustrates that, in such a situation, the likelihood of establishment is maximised when the community size $N_\mathrm{n}$ is small, i.e. when consanguinity increases the chance that mutant homozygotes may be produced.

# CHAPTER 5 CONCLUSION

The Bayesian network presented in Section 3.3.3 amalgamates a number of disparate inputs of various types into an estimate of the presence of CF in an individual. While an attempt was made to incorporate most of the commonly used diagnostic inputs into the model, it is of course by no means complete. The network can easily be extended, if desired, although it would probably make more sense to rather focus on improving the joint probability distribution functions between existing nodes, as noted in section 3.4.2. Nonetheless, despite the questionable nature of some of the underlying assumptions, the model does compute plausible probabilities, at least in some boundary cases where reasonably accurate known statistics exist.

The model is not presented as a practical diagnostic tool, mainly due to the same limiting factors identified in the literature study presented in section 2.2.5. However, interacting with the model may be of educational value to health professionals and students, because of the sometimes non-intuitive outcomes when evidence is combined, and it could help to guide decisions regarding treatment and/or additional testing, by indicating where the most value is likely to be added.

The population simulation tool described in Chapter 4 was demonstrated to replicate various analytically predicted results, as well as several observations, while only requiring a single fairly constrained estimate – the effective neighbourhood size – which is based on sociological data for human populations. As far as could be ascertained this is the first time that this input has been proposed to augment population genetics models.

The described verification results add credibility to the results that are generated by the model, starting with a plausible estimate for the heterozygous selective advantage conferred by CF and similar deleterious recessive autosomal monogenic conditions, without requiring specific knowledge regarding the mechanism by which the selective advantage is achieved.

It is also shown that the generally accepted heterozygous carrier prevalence levels of deleterious alleles in various countries and population groups are actually calculated values, using specious implicit assumptions of panmixia and equilibrium. These carrier prevalence values should be adjusted downwards by a significant margin, implying that there are far fewer human carriers of recessive deleterious alleles than previously thought.

## 5.1 SUGGESTIONS FOR FUTURE STUDY

### 5.1.1 Diagnosis

The Bayesian network for diagnostic purposes in its current form is limited by the availability of observational data, reducing its utility to mostly tutorial value. Even though the model is shown to be remarkably robust to noise, more accurate inputs will directly improve its performance. A large number of other inputs can also be added, including but not limited to those listed by O’Sullivan and Freedman [1] and Farrell *et al* [6].

The same lack of data stymied the ambition of incorporating and studying the temporal feature inherent in health care [159], which, as stated in section 3.3.2, seems to be a promising area for further exploration, because the longitudinal element inherent in cohort studies may possibly be incorporated by the time-dependence of dynamic Bayesian models, of which hidden Markov models are a subset [160].

Furthermore attention may be given to the identification of latent variables which may affect and/or connect known parameters in the model, and potentially interfere with causal interpretation.

### 5.1.2 Population Models

There are a large number of CFTR alleles associated with cystic fibrosis, each with its own heterozygous and homozygous selection coefficients. Currently only one variant is modelled at a time. As in the case of the diagnostic network, one of the primary challenges would be to obtain accurate observational data, especially for the huge number of potential allele combinations that multiple variants would imply. If this could be done, though, it may also be feasible to not only model the establishment, dissemination, and equilibrium behaviour, but also competition between variants.

The population model currently simulates a single homogeneous population of constant size, without any boundaries or discontinuities. It is well known that geographic features and changing population sizes affect speciation as well as establishment rates - the model can be extended to include these attributes, which should improve its credibility significantly.

# REFERENCES

[1] B. P. O'Sullivan and S. D. Freedman, "Cystic fibrosis," *Lancet*, vol. 373, no. 9678, pp. 1891-1904, May 2009.

[2] I. C. Withrock *et al*, "Genetic diseases conferring resistance to infectious diseases," *Genes Dis.*, vol. 2, no. 3, pp. 247-254, September 2015.

[3] F. B. Piel, A. P. Patil, R. E. Howes, O. A. Nyangiri, P. W. Gething, T. N. Williams, D. J. Weatherall and S. I. Hay, "Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis," *Nat. Commun.*, vol. 1, no. 8, November 2010.

[4] P. W. Hedrick, "Population genetics of malaria resistance in humans," *Heredity*, vol. 107, no. 4, pp. 283-304, March 2011.

[5] B. A. Poku, A. L. Caress and S. Kirk, "Adolescents' experiences of living with sickle cell disease: An integrative narrative review of the literature," *Int. J. Nurs. Stud.*, vol. 80, pp. 20-28, April 2018.

[6] P. M. Farrell *et al*, "Diagnosis of cystic fibrosis: consensus guidelines from the Cystic Fibrosis Foundation," *J. Pediatr.*, vol. 181, no. Supplement, pp. S4-S15, February 2017.

[7] A. Tluczek, K. M. Orland and L. Cavanagh, "Psychosocial consequences of false-positive newborn screens for cystic fibrosis," *Qual. Health Res.*, vol. 21, no. 2, pp. 174-186, September 2010.

[8] J. van Rensburg, M. Alessandrini, C. Stewart and M. S. Pepper, "Cystic fibrosis in South Africa: a changing diagnostic paradigm," *S. Afr. Med. J.,* vol. 108, no. 8, pp. 624-628, August 2018.

[9] S. Wright, "Evolution in Mendelian populations," *Bull. Math. Biol.,* vol. 52, no. 1-2, pp. 241-295, January 1990.

[10] L. Nunney, "The effect of neighborhood size on effective population size in theory and in practice," *Heredity,* vol. 117, no. 4, pp. 224-232, August 2016.

[11] M. Kimura and J. F. Crow, "The number of alleles that can be maintained in a finite population," *Genetics,* vol. 49, pp. 725-738, April 1964.

[12] Z. Patwa and L. M. Wahl, "The fixation probability of beneficial mutations," *J. R. Soc. Interface,* vol. 5, no. 28, pp. 1279-1289, July 2008.

[13] R. I. M. Dunbar, "Neocortex size as a constraint on group size in primates," *J. Hum. Evol.,* vol. 22, no. 6, pp. 469-493, June 1992.

[14] J. Lehmann, A. H. Korstjens and R. I. M. Dunbar, "Group size, grooming and social cohesion in primates," *Anim. Behav.,* vol. 74, no. 6, pp. 1617-1629, December 2007.

[15] J. W. Viljoen, J. P. de Villiers, A. J. van Zyl, M. Mezzavilla and M. Pepper, "Establishment and equilibrium levels of deleterious mutations in large populations," *Nature Scientific Reports,* July 2019.

[16] A. C. Doyle, *The great Keinplatz experiment and other tales of twilight and the unseen*, Ed. public domain, Project Gutenberg, June 2010.

[17] D. T. Durack, "The weight of medical knowledge," *N. Engl. J. Med.*, vol. 298, no. 14, pp. 773-775, April 1978.

[18] D. J. d. S. Price, *Little science, big science - and beyond*, New York: Columbia University Press, September 1986.

[19] P. G. Ramsey, J. D. Carline, T. S. Inui, E. B. Larson, J. P. Logerfo, J. J. Norcini and M. D. Wenrich, "Changes over time in the knowledge base of practicing internists," *J. Am. Med. Assoc.*, vol. 266, no. 8, pp. 1103-1107, August 1991.

[20] J. Boswell, *The Life of Samuel Johnson, LL.D. vol. 2*, London: Charles Dilly, 1791.

[21] D. E. Newman-Toker and P. J. Pronovost, "Diagnostic errors the next frontier for patient safety," *J. Am. Med. Assoc.*, vol. 301, no. 10, pp. 1060-1062, March 2009.

[22] R. A. Miller, "Computer-assisted diagnostic decision support: history, challenges, and possible paths forward," *Adv. Health Sci. Educ. Theory Pract.*, vol. 14, no. 1 suppl., pp. 89-106, September 2009.

[23] R. A. Miller, "Why the standard view is standard: People, not machines, understand patients' problems," *J. Med. Philos. (UK)*, vol. 15, no. 6, pp. 581-591, December 1990.

[24] R. S. Ledley and L. B. Lusted, "Reasoning foundations of medical diagnosis," *Science*, vol. 130, no. 3366, pp. 9-21, July 1959.

[25] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," *Psychol. Rev.*, vol. 65, no. 6, pp. 386-408, November 1958.

[26] D. L. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proc. IEEE,* vol. 85, no. 1, pp. 6-23, January 1997.

[27] M. M. Kokar, J. A. Tomasik and J. Weyman, "Formalizing classes of information fusion systems," *Inf. Fusion,* vol. 5, no. 3, pp. 189-202, September 2004.

[28] E. H. Shortliffe, S. G. Axline, B. G. Buchanan, T. C. Merigan and S. N. Cohen, "An artificial intelligence program to advise physicians regarding antimicrobial therapy," *Comput. Biomed. Res.,* vol. 6, no. 6, pp. 544-560, December 1973.

[29] E. H. Shortliffe, F. S. Rhame and S. G. Axline, "Mycin, a computer program providing antimicrobial therapy recommendations," *Clin. Res.,* vol. 23, no. 2, August 1975.

[30] T. Bayes and R. Price, "An essay towards solving a problem in the doctrine of chances," *Philos. Trans. R. Soc. Lond.,* vol. 53, no. 2, pp. 370-418, December 1763.

[31] J. B. Michel *et al*, "Quantitative analysis of culture using millions of digitized books," *Science,* vol. 331, no. 6014, pp. 176-182, January 2011.

[32] H. Jeffreys, *Scientific Inference,* Ed. 3, London: Cambridge University Press, 1973.

[33] P. M. Farrell, "The prevalence of cystic fibrosis in the European Union," *J. Cyst. Fibros.,* vol. 7, no. 5, pp. 450-453, April 2008.

[34] A. P. Dempster, "A generalization of Bayesian inference," *J. R. Stat. Soc.,* vol. 3, no. 2, pp. 205-247, February 1968.

[35] G. Shafer, *A Mathematical Theory of Evidence*, Princeton: Princeton University Press, April 1976.

[36] G. Shafer, "Perspectives on the theory and practice of belief functions," *Int. J. Approximate Reasoning,* vol. 4, no. 5-6, pp. 323-362, October 1990.

[37] J. Kohlas and P. Monney, *A mathematical theory of hints. An approach to the Dempster-Shafer theory of evidence*, Berlin: Springer-Verlag, July 1995.

[38] A. Jøsang and R. Hankin, "Interpretation and fusion of hyper opinions in subjective logic," in *15th International Conference on Information Fusion, FUSION 2012*, Singapore, pp. 1225-1232, September 2012.

[39] G. Hooper, "A calculation of the credibility of human testimony," *Philos. Trans. R. Soc.*, vol. 21, no. 1, pp. 359-365, 1699.

[40] L. A. Zadeh, "Simple view of the Dempster-Shafer theory of evidence and its implication for the rule of combination," *AI Mag.,* vol. 7, no. 2, pp. 85-90, June 1986.

[41] J. Pearl, "Reasoning with belief functions: an analysis of compatibility," *Int. J. Approximate Reasoning,* vol. 4, no. 5-6, pp. 363-389, September 1990.

[42] A. Gelman, "The boxer, the wrestler, and the coin flip: a paradox of robust Bayesian inference and belief functions," *Am. Stat.,* vol. 60, no. 2, pp. 146-150, May 2006.

[43] J. Dezert, P. Wang and A. Tchamova, "On the validity of Dempster-Shafer theory," in *15th International Conference on Information Fusion, FUSION 2012*, Singapore, pp. 655-660, September 2012.

[44] D. E. Rumelhart, B. Widrow and M. A. Lehr, "The basic ideas in neural networks," *Commun. ACM*, vol. 37, no. 3, pp. 87-92, March 1994.

[45] K. L. Priddy and P. E. Keller, *Artificial Neural Networks : An Introduction*, Ed. 1, Bellingham, Washington: SPIE-The International Society for Optical Engineering, 2005.

[46] P. Mazzoni, R. A. Andersen and M. I. Jordan, "A more biologically plausible learning rule than backpropagation applied to a network model of cortical area 7a," *Cereb. Cortex*, vol. 1, no. 4, pp. 293-307, July 1991.

[47] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Math. Control Signal.*, vol. 2, no. 4, pp. 303-314, December 1989.

[48] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol.*, vol. 160, no. 1, pp. 106-154, January 1962.

[49] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *J. Physiol.*, vol. 195, no. 1, pp. 215-243, March 1968.

[50] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biol. Cybern.*, vol. 36, no. 4, pp. 193-202, April 1980.

[51] J. J. Weng, N. Ahuja and T. S. Huang, "Learning recognition and segmentation of 3-D objects from 2-D images," in *1993 IEEE 4th International Conference on Computer Vision*, pp. 121-127, May 1993.

[52] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artif. Intell. Med.*, vol. 23, no. 1, pp. 89-109, April 2001.

[53] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273-297, September 1995.

[54] G. Orrù, W. Pettersson-Yeo, A. F. Marquand, G. Sartori and A. Mechelli, "Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: A critical review," *Neurosci. Biobehav. Rev.*, vol. 36, no. 4, pp. 1140-1152, April 2012.

[55] S. Klöppel *et al*, "Automatic classification of MR scans in Alzheimer's disease," *Brain*, vol. 131, no. 3, pp. 681-689, March 2008.

[56] H. Peng, F. Long and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226-1238, August 2005.

[57] H.-L. Chen, B. Yang, G. Wang, S.-J. Wang, J. Liu and D.-Y. Liu, "Support vector machine based diagnostic system for breast cancer using swarm intelligence," *J. Med. Syst.*, vol. 36, no. 4, pp. 2505-2519, August 2012.

[58] F. Nigsch, A. Bender, B. van Buuren, J. Tissen, E. Nigsch and J. B. O. Mitchell, "Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization," *J. Chem. Inf. Model.*, vol. 46, no. 6, pp. 2412-2422, November 2006.

[59] L. A. Zadeh, "Fuzzy sets," *Inf. Control,* vol. 8, no. 3, pp. 338-353, June 1965.

[60] J. A. Goguen, "L-fuzzy sets," *J. Math. Anal. Appl.,* vol. 18, no. 1, pp. 145-174, April 1967.

[61] K. Hiramatsu, K. Kabasawa and S. Kaihara, "Application of the fuzzy logic to medical diagnosis," *Iyodenshi. to. Seitai Kogaku,* vol. 12, no. 3, pp. 148-155, June 1974.

[62] H. Wechsler, "A fuzzy approach to medical diagnosis," *Int. J. Biomed. Comput.,* vol. 7, no. 3, pp. 191-203, July 1976.

[63] W. van Leekwijck and E. E. Kerre, "Defuzzification: criteria and classification," *Fuzzy Sets Syst.,* vol. 108, no. 2, pp. 159-178, December 1999.

[64] H. R. Warner, A. F. Toronto, L. G. Veasey and R. Stephenson, "A mathematical approach to medical diagnosis: application to congenital heart disease," *J. Am. Med. Assoc.,* vol. 177, no. 3, pp. 177-183, July 1961.

[65] G. A. Gorry and G. O. Barnett, "Experience with a model of sequential diagnosis," *Comput. Biomed. Res.,* vol. 1, no. 5, pp. 490-507, May 1968.

[66] G. A. Gorry, "Strategies for computer-aided diagnosis," *Math. Biosci.,* vol. 2, no. 3-4, pp. 293-318, May 1968.

[67] R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum and J. Lederberg, "DENDRAL: A case study of the first expert system for scientific hypothesis formation," *Artif. Intell.,* vol. 61, no. 2, pp. 209-261, June 1993.

[68] D. E. Heckerman and E. H. Shortliffe, "From certainty factors to belief networks," *Artif. Intell. Med.,* vol. 4, no. 1, pp. 35-52, February 1992.

[69] R. Miller, F. E. Masarie and J. D. Myers, "Quick Medical Reference (QMR) for diagnostic assistance," *M.D. Comput.,* vol. 3, no. 5, pp. 34-48, September 1986.

[70] M. Shwe and G. F. Cooper, "An empirical analysis of likelihood-weighting simulation on a large, multiply connected medical belief network," *Comput. Biomed. Res.,* vol. 24, no. 5, pp. 453-475, October 1991.

[71] G. F. Cooper, "The computational complexity of probabilistic inference using Bayesian belief networks," *Artif. Intell.,* vol. 42, no. 2, pp. 393-405, March 1990.

[72] D. A. B. Lindberg, L. R. Rowland, C. R. Buch, W. F. Morse and S. S. Morse, "CONSIDER: A computer program for medical instruction," in *Proceedings of the Ninth IBM Medical Symposium*, Burlington, Vermont, pp. 59-61, October 1968.

[73] S. G. Pauker, G. A. Gorry, J. P. Kassirer and W. B. Schwartz, "Towards the simulation of clinical cognition. Taking a present illness by computer," *Am. J. Med.,* vol. 60, no. 7, pp. 981-996, June 1976.

[74] S. M. Weiss, C. A. Kulikowski and A. Safir, "Glaucoma consultation by computer," *Comput. Biol. Med.,* vol. 8, no. 1, pp. 25-40, January 1978.

[75] S. M. Weiss, K. B. Kern, C. A. Kulikowski and W. Pincus, "Interactive system for the design of classifiers in diagnostic applications," in *Proc. Int. Conf. Cybern. Soc.*, Kyoto, Jpn, pp. 58-62, November 1978.

[76] M. S. Blois, M. S. Tuttle and D. D. Sherertz, "RECONSIDER: A program for generating differential diagnoses," in *Proceedings - Annual Symposium on Computer Applications in Medical Care*, Washington DC, USA, pp. 219-221, November 1981.

[77] S. J. Nelson, M. S. Blois, M. S. Tuttle, M. Erlbaum, P. Harrison, H. Kim, B. Winkelmann and D. Yamashita, "Evaluating RECONSIDER - A computer program for diagnostic prompting," *J. Med. Syst.*, vol. 9, no. 5-6, pp. 379-388, December 1985.

[78] K. Adlassnig and G. Kolarz, "CADIAG-2: Computer-assisted medical diagnosis using fuzzy subsets," in *Approximate Reasoning in Decision Analysis*, pp. 219-247, Amsterdam, Neth.: North-Holland Publ. Co., January 1982.

[79] R. A. Miller, H. E. Pople and J. D. Myers, "Internist-I, an experimental computer-based diagnostic consultant for general internal medicine," *N. Engl. J. Med.*, vol. 307, no. 8, pp. 468-476, August 1982.

[80] L. W. Gaston, D. A. B. Lindberg, A. D. Vanker and L. Kingsland III, "AI/COAG, a knowledge-based surrogate for the human hemostasis expert," *Mo. Med.*, vol. 80, no. 4, pp. 185-188, April 1983.

[81] P. Politakis and S. M. Weiss, "Using empirical analysis to refine expert system knowledge bases," *Artif. Intell.*, vol. 22, no. 1, pp. 23-48, January 1984.

[82] E. J. Horvitz, D. E. Heckerman, B. N. Nathwani and L. M. Fagan, "Diagnostic strategies in the hypothesis-directed PATHFINDER system," in *First Conference on Artificial Intelligence Applications*, Denver, CO, USA, pp. 630-636, December 1984.

[83] G. O. Barnett, J. J. Cimino, J. A. Hupp and E. P. Hoffer, "DXplain. An evolving diagnostic decision-support system," *J. Am. Med. Assoc.*, vol. 258, no. 1, pp. 67-74, July 1987.

[84] H. R. Warner, P. Haug, O. Bouhaddou, M. Lincoln, H. Warner Jr, D. Sorenson, J. W. Williamson and C. Fan, "ILIAD as an expert consultant to teach differential diagnosis," in *Proceedings - Annual Symposium on Computer Applications in Medical Care*, Washington DC, USA, pp. 371-376, November 1988.

[85] Y. Yoon, L. L. Peterson and P. R. Bergstresser, "DESKNET: The dermatology expert system with knowledge-based network," *Neural Networks*, vol. 1, no. 1 suppl., p. 477, September 1988.

[86] H. S. Waxman and W. E. Worley, "Computer-assisted adult medical diagnosis: Subject review and evaluation of a new microcomputer-based system," *Medicine*, vol. 69, no. 3, pp. 125-136, May 1990.

[87] J. R. Clarke, M. Niv, B. L. Webber, K. Fisherkeller and B. L. Ryack, "TraumAID: A decision aid for managing trauma at various levels of resources," in *Proceedings: Thirteenth Annual Symposium on Computer Applications in Medical Care (SCAMC-13)*, Washington DC, USA, pp. 1005-1006, November 1989.

[88] A. Geissbuhler and R. A. Miller, "Clinical application of the UMLS in a computerized order entry and decision-support system," in *Proceedings of the Annual AMIA Symposium*, pp. 320-324, February 1998.

[89] B. Kanagaratnam, S. Lavelie and R. Comerford, "FTO1/362: Using neural nets in medical decision making," *J. Med. Internet Res.*, vol. 1, no. 1 suppl., pp. 41-42, September 1999.

[90] L. Godo, R. L. de Mántaras, J. Puyol-Gruart and C. Sierra, "Renoir, Pneumon-IA and Terap-IA: three medical applications based on fuzzy logic," *Artif. Intell. Med.,* vol. 21, no. 1, pp. 153-162, January 2001.

[91] N. Belacel and M. R. Boulassel, "Multicriteria fuzzy classification procedure PROCFTN: methodology and medical application," *Fuzzy Sets Syst.,* vol. 141, no. 2, pp. 203-217, January 2004.

[92] B. Wemmenhove, J. M. Mooij, W. Wiegerinck, M. Leisink, H. J. Kappen and J. P. Neijt, "Inference in the PROMEDAS medical expert system," in *11th Conference on Artificial Intelligence in Medicine*, Amsterdam, Neth., pp. 456-460, July 2007.

[93] R. Carvalho, R. Isola and A. K. Tripathy, "MediQuery - An automated decision support system," in *Proceedings - 24th International Symposium on Computer-Based Medical Systems*, Bristol, UK, June 2011.

[94] D. Ferrucci, A. Levas, S. Bagchi, D. Gondek and E. T. Mueller, "Watson: Beyond Jeopardy!," *Artif. Intell.,* vol. 199-200, pp. 93-105, June 2013.

[95] D. Ferrucci *et al*, "Building Watson: An overview of the deepQA project," *AI Mag.,* vol. 31, no. 3, pp. 59-79, September 2011.

[96] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola and L. K. Saul, "Introduction to variational methods for graphical models," *Mach. Learn.,* vol. 37, no. 2, pp. 183-233, November 1999.

[97] M. I. Jordan, "Graphical models," *Stat. Sci.,* vol. 19, no. 1, pp. 140-155, February 2004.

[98] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*, Ed. 2, San Francisco: Morgan Kaufmann, 1986.

[99] T. S. Jaakkola and M. I. Jordan, "Variational probabilistic inference and the QMR-DT network," *J. Artif. Intell. Res.*, vol. 10, pp. 291-322, January 1999.

[100] C. S. Jensen, U. Kjærulff and A. Kong, "Blocking Gibbs sampling in very large probabilistic expert systems," *Int. J. Hum. Comput. Stud.*, vol. 42, no. 6, pp. 647-666, June 1995.

[101] R. W. White and E. Horvitz, "Cyberchondria: studies of the escalation of medical concerns in web search," *ACM Trans. Inf. Syst.*, vol. 27, no. 4, November 2009.

[102] J. K. Jerome, *Three men in a boat (to say nothing of the dog)*, Bristol & London: J.W. Arrowsmith & Simpkin, 1889.

[103] E. Strickland, "Watson goes to med school: IBM's AI program mastered 'Jeopardy!' next up, oncology," *IEEE Spectrum*, vol. 50, no. 1, pp. 42-45, January 2013.

[104] S. Kaisler, F. Armour, J. A. Espinosa and W. Money, "Big data: Issues and challenges moving forward," in *Proceedings of the Annual Hawaii International Conference on System Sciences*, Wailea, Maui, HI, pp. 995-1004, January 2013.

[105] T. Tanaka, "Big data application technology: an overview," *IEEJ Trans. Electron. Inf. Syst.*, vol. 133, no. 3, pp. 550-553, March 2013.

[106] D. S. Ting *et al*, "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *J. Am. Med. Assoc.,* vol. 318, no. 22, pp. 2211-2223, October 2017.

[107] B. E. Bejnordi *et al*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *J. Am. Med. Assoc.,* vol. 318, no. 22, pp. 2199-2210, December 2017.

[108] V. Gulshan *et al*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *J. Am. Med. Assoc.,* vol. 316, no. 22, pp. 2402-2410, December 2016.

[109] A. M. Turing, "Computing machinery and intelligence," *Mind,* vol. 49, no. 236, pp. 433-460, October 1950.

[110] B. Stanberry, "Telemedicine: barriers and opportunities in the 21st century," *J. Intern. Med.,* vol. 247, no. 6, pp. 615-628, June 2000.

[111] G. B. Pier *et al*, "Salmonella typhi uses CFTR to enter intestinal epithelial cells," *Nature,* vol. 393, no. 6680, pp. 79-82, May 1998.

[112] S. A. Schroeder, D. M. Gaughan and M. Swift, "Protection against bronchial asthma by CFTR Δf508 mutation: a heterozygote advantage in cystic fibrosis," *Nat. Med.,* vol. 1, no. 7, pp. 703-705, July 1995.

[113] P. C. Doherty and R. M. Zinkernagel, "Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex," *Nature,* vol. 256, no. 5512, pp. 50-52, July 1975.

[114] J. B. S. Haldane, "A mathematical theory of natural and artificial selection, part V: Selection and mutation," *Math. Proc. Camb. Philos. Soc.*, vol. 23, no. 7, pp. 838-844, July 1927.

[115] M. Kimura, "On the probability of fixation of mutant genes in a population," *Genetics*, vol. 47, pp. 713-719, June 1962.

[116] S. Wright, "Isolation by distance under diverse systems of mating," *Genetics*, vol. 31, pp. 39-59, January 1946.

[117] T. Dobzhansky and S. Wright, "Genetics of natural populations. V. Relations between mutation rate and accumulation of lethals in populations of Drosophila pseudoobscura," *Genetics*, vol. 26, pp. 23-51, January 1941.

[118] T. Dobzhansky and S. Wright, "Genetics of natural populations. X. Dispersion rates in Drosophila pseudoobscura," *Genetics*, vol. 28, pp. 304-340, July 1943.

[119] S. Wright, "Isolation by distance," *Genetics*, vol. 28, pp. 114-138, March 1943.

[120] S. Wright, T. Dobzhansky and W. Hovanitz, "Genetics of natural populations. VII. The allelism of lethals in the third chromosome of Drosophila pseudoobscura," *Genetics*, vol. 27, pp. 363-394, July 1942.

[121] R. D. H. Barrett, L. K. M'Gonigle and S. P. Otto, "The distribution of beneficial mutant effects under strong selection," *Genetics*, vol. 174, no. 4, pp. 2071-2079, 2006.

[122] M. Kimura, "Some problems of stochastic processes in genetics," *Ann. Math. Stat.*, vol. 28, pp. 882–901, 1957.

[123] M. Kimura, "Diffusion models in population genetics," *J, Appl. Probab.*, vol. 1, pp. 177-232, 1964.

[124] N. Firdous, S. Gibbons and B. Modell, "Falling prevalence of beta-thalassaemia and eradication of malaria in the Maldives," *J. Community Genet.*, vol. 2, no. 3, pp. 173-189, September 2011.

[125] A. G. Motulsky, "Jewish diseases and origins," *Nat. Genet.*, vol. 9, no. 2, pp. 99-101, February 1995.

[126] J. H. Pearn, "The gene frequency of acute Werdnig-Hoffmann disease (SMA type 1). A total population survey in North-East England," *J. Med. Genet.*, vol. 10, no. 3, pp. 260-265, September 1973.

[127] J. R. Butterworth *et al*, "The role of hemochromatosis susceptibility gene mutations in protecting against iron deficiency in celiac disease," *Gastroenterology*, vol. 123, no. 2, pp. 444-449, August 2002.

[128] E. D. Weinberg, "Survival advantage of the hemochromatosis C282Y mutation," *Perspect. Biol. Med.*, vol. 51, no. 1, pp. 98-102, December 2008.

[129] A. M. Tuli, R. K. Valenzuela, E. Kamugisha and M. H. Brilliant, "Albinism and disease causing pathogens in Tanzania: Are alleles that are associated with OCA2 being maintained by balancing selection?," *Med. Hypotheses*, vol. 79, no. 6, pp. 875-878, December 2012.

[130] J. E. A. Common, W. L. Di, D. Davies and D. P. Kelsell, "Further evidence for heterozygote advantage of GJB2 deafness mutations: A link with cell survival," *J. Med. Genet.*, vol. 41, no. 7, pp. 573-575, July 2004.

[131] S. J. Moore *et al*, "Clinical and genetic epidemiology of Bardet-Biedl syndrome in Newfoundland: A 22-year prospective, population-based, cohort study," *Am. J. Med. Genet.,* vol. 132 A, no. 4, pp. 352-360, February 2005.

[132] M. Pandolfo and A. Pastore, "The pathogenesis of Friedreich ataxia and the structure and function of frataxin," *J. Neurol.,* vol. 256, no. 1 suppl., pp. 9-17, March 2009.

[133] M. A. Sadat *et al*, "Glycosylation, hypogammaglobulinemia, and resistance to viral infections," *N. Engl. J. Med.,* vol. 370, no. 17, pp. 1615-1625, April 2014.

[134] S. Wright and T. Dobzhansky, "Genetics of natural populations. XII. Experimental reproduction of some of the changes caused by natural selection in certain populations of Drosophila pseudoobscura," *Genetics,* vol. 31, pp. 125-156, March 1946.

[135] T. Dobzhansky and S. Wright, "Genetics of natural populations. XV. Rate of diffusion of a mutant gene through a population of Drosophila pseudoobscura," *Genetics,* vol. 32, no. 3, pp. 303-324, May 1947.

[136] N. P. Dubinin, "On lethal mutations in natural populations," *Genetics,* vol. 31, pp. 21-38, January 1946.

[137] T. Greenhalgh, "How to read a paper: Getting your bearings (deciding what the paper is about)," *Br. Med. J.,* vol. 315, no. 7102, pp. 243-246, July 1997.

[138] M. Angell, "The Nazi hypothermia experiments and unethical research today," *N. Engl. J. Med.,* vol. 322, no. 20, pp. 1462-1464, May 1990.

[139] J. B. Nie, "The United States cover-up of Japanese wartime medical atrocities: Complicity committed in the national interest and two proposals for contemporary action," *Am. J. Bioethics,* vol. 6, no. 3, pp. 21-33, July 2006.

[140] R. L. Berger, "Nazi science — the Dachau hypothermia experiments," *N. Engl. J. Med.,* vol. 322, no. 20, pp. 1435-1440, May 1990.

[141] H. K. Beecher, "Ethics and clinical research," *N. Engl. J. Med.,* vol. 274, no. 24, pp. 1354-1360, June 1966.

[142] J. Pearl, "Causal inference in the health sciences: a conceptual introduction," *Health Serv. Outcomes Res. Method.,* vol. 2, no. 3-4, pp. 189-220, December 2001.

[143] R. Doll and A. B. Hill, "Smoking and carcinoma of the lung preliminary report," *Br. Med. J.,* vol. 2, no. 4682, pp. 739-748, September 1950.

[144] M. Ezzati, S. J. Henley, A. D. Lopez and M. J. Thun, "Role of smoking in global and regional cancer epidemiology: current patterns and data needs," *Int. J. Cancer,* vol. 116, no. 6, pp. 963-971, October 2005.

[145] J. Cornfield, "A method of estimating comparative rates from clinical data: applications to cancer of the lung, breast, and cervix," *J. Natl. Cancer Inst.,* vol. 11, no. 6, pp. 1269-1275, June 1951.

[146] O. Miettinen, "Estimability and estimation in case-referent studies," *Am. J. Epidemiol.,* vol. 103, no. 2, pp. 226-235, February 1976.

[147] H. F. Dorn, "Some problems arising in prospective and retrospective studies of the etiology of disease," *N. Engl. J. Med.,* vol. 261, pp. 571-579, September 1959.

[148] N. Mantel and W. Haenszel, "Statistical aspects of the analysis of data from retrospective studies of disease," *J. Natl. Cancer Inst.,* vol. 22, no. 4, pp. 719-748, April 1959.

[149] P. Cole, "The evolving case-control study," *J. Chronic Dis.,* vol. 32, no. 1-2, pp. 15-27, January 1979.

[150] J. Pearl, "Causal inference in statistics: an overview," *Stat. Surv.,* vol. 3, pp. 96-146, January 2009.

[151] A. S. Rodin, G. Gogoshin, A. Litvinenko and E. Boerwinkle, "Exploring genetic epidemiology data with Bayesian networks," in *Handbook of Statistics vol. 28*, pp. 479-510, Amsterdam, Neth.: Elsevier, January 2012.

[152] P. Kraft, Y. C. Yen, D. O. Stram, J. Morrison and W. J. Gauderman, "Exploiting gene-environment interaction to detect genetic associations," *Hum. Hered.,* vol. 63, no. 2, pp. 111-119, February 2007.

[153] C. Su, A. Andrew, M. R. Karagas and M. E. Borsuk, "Using Bayesian networks to discover relations between genes, environment, and disease," *BioData Min.,* vol. 6, no. 1, March 2013.

[154] J. H. Moore, "The ubiquitous nature of epistasis in determining susceptibility to common human diseases," *Hum. Hered.,* vol. 56, no. 1-3, pp. 73-82, November 2003.

[155] J. H. Moore, J. C. Gilbert, C. T. Tsai, F. T. Chiang, T. Holden, N. Barney and B. C. White, "A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility," *J. Theor. Biol.,* vol. 241, no. 2, pp. 252-261, July 2006.

[156] M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl and J. H. Moore, "Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer," *Am. J. Hum. Genet.,* vol. 69, no. 1, pp. 138-147, July 2001.

[157] M. D. Ritchie, L. W. Hahn and J. H. Moore, "Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity," *Genet. Epidemiol.,* vol. 24, no. 2, pp. 150-157, February 2003.

[158] N. E. Breslow, "Statistics in epidemiology: the case-control study," *J. Am. Stat. Assoc.,* vol. 91, no. 433, pp. 14-28, March 1996.

[159] S. Kleinberg and G. Hripcsak, "A review of causal inference for biomedical informatics," *J. Biomed. Informatics,* vol. 44, no. 6, pp. 1102-1112, December 2011.

[160] G. Theocharous, K. Murphy and L. P. Kaelbling, "Representing hierarchical POMDPs as DBNs for multi-scale robot localization," in *Proceedings - IEEE International Conference on Robotics and Automation,* New Orleans, LA, pp. 1045-1051, May 2004.

[161] S. M. Moskowitz, J. F. Chmiel, D. L. Sternen, E. Cheng, R. L. Gibson, S. G. Marshall and G. R. Cutting, "Clinical practice and genetic counseling for cystic fibrosis and CFTR-related disorders," *Genet. Med.,* vol. 10, no. 12, pp. 851-868, December 2008.

[162] J. Massie and M. B. Delatycki, "Cystic fibrosis carrier screening," *Paediatr. Respir. Rev.,* vol. 14, no. 4, pp. 270-275, December 2013.

[163] Statistics South Africa, *Mid-year population estimates 2018,* Pretoria: Stats SA, July 2018.

[164] K. Lowton and J. Gabe, "Life on a slippery slope: Perceptions of health in adults with cystic fibrosis," *Sociol. Health Illn.,* vol. 25, no. 4, pp. 289-319, May 2003.

[165] P. M. Farrell, S. Joffe, L. Foley, G. J. Canny, P. Mayne and M. Rosenberg, "Diagnosis of cystic fibrosis in the Republic of Ireland: epidemiology and costs," *Ir. Med. J.,* vol. 100, no. 8, September 2007.

[166] A. Orenti *et al, ECFSPR Annual Report 2016,* Denmark: European Cystic Fibrosis Society, 2018.

[167] D. M. Swallow, "Genetics of lactase persistence and lactose intolerance," *Annu. Rev. Genet.,* vol. 37, pp. 197-219, July 2003.

[168] P. Gerbault, A. Liebert, Y. Itan, A. Powell, M. Currat, J. Burger, D. M. Swallow and M. G. Thomas, "Evolution of lactase persistence: An example of human niche construction," *Philos. Trans. R. Soc. B Biol. Sci.,* vol. 366, no. 1566, pp. 863-877, February 2011.

[169] G. D. Smith, D. A. Lawlor, N. J. Timpson, J. Baban, M. Kiessling, I. N. M. Day and S. Ebrahim, "Lactase persistence-related genetic variant: population substructure and health outcomes," *Eur. J. Hum. Genet.,* vol. 17, no. 3, pp. 357-367, March 2009.

[170] T. Bersaglieri *et al,* "Genetic signatures of strong recent positive selection at the lactase gene," *Am. J. Hum. Genet.,* vol. 74, no. 6, pp. 1111-1120, April 2004.

[171] J. N. Fenner, "Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies," *Am. J. Phys. Anthropol.,* vol. 128, no. 2, pp. 415-423, March 2005.

[172] A. Light and P. J. Bartlein, "The end of the rainbow? Color schemes for improved data graphics," *Eos,* vol. 85, no. 40, October 2004.

[173] M. Lek *et al,* "Analysis of protein-coding genetic variation in 60,706 humans," *Nature,* vol. 536, no. 7616, pp. 285-291, 2016.

[174] K. Karczewski *et al*, "Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes," *bioRxiv,* 2019.

[175] D. P. Doolittle, *Population Genetics : Basic Principles*, Berlin: Springer-Verlag, 1987.

[176] P. M. Altrock, A. Traulsen and F. A. Reed, "Stability properties of underdominance in finite subdivided populations," *PLoS Comput. Biol.,* vol. 7, no. 11, November 2011.

[177] J. B. S. Haldane, "Selection against heterozygosis in man," *Ann. Hum. Genet.,* vol. 11, no. 1, pp. 333-340, January 1941.

[178] K. Magori and F. Gould, "Genetically engineered underdominance for manipulation of pest populations: a deterministic model," *Genetics,* vol. 172, no. 4, pp. 2613-2620, April 2006.

# ADDENDUM A DERIVATION OF EQUILIBRIUM PREVALENCE LEVELS

In special cases analytical results, with which one can compare the simulated results, may be derived for this model. This section presents, in the case of a large neighbourhood size, a heuristic derivation of carrier prevalence which results in a simple relationship between the long-term prevalence of the mutation and the heterozygous advantage. The (random) state of the individual at grid point ($i$,$j$) in the population array can be described as a stochastic process $(X_n^{i,j})_{n=0}^{\infty}$, where $n$ denotes the number of generations from the beginning. We use $w$ for the wild-type, $e$ for the heterozygous and $m$ for the homozygous states respectively.

The states of different individuals are distributed identically, in other words they are statistically indistinguishable because the development of a specific individual depends only on its neighbours and not the absolute position ($i$,$j$). (This is due to the fact that the population is closed upon itself, so that there are no edges or other position-specific effects.) Hence, we can drop the superscripts and denote the state of a generic individual at the $n^{th}$ generation by $X_n$.

Our concern is with the temporal evolution of the quantity $p_n := P(X_n = e)$; that is, the probability that a generic individual is a heterozygous carrier of the mutation. Additionally, we would like to derive the existence and value of the equilibrium level $p$ where $p_n = p$ implies $p_{n+1} = p$. For large populations, the value of $p$ indicates the prevalence of the mutation; that is, the percentage level at which the heterozygous advantage balances out the homozygous disadvantage, on average.

We assume, for ease of display, that the homozygous *dis*advantage is 100% (i.e. $s_{hom} = -1$) so that homozygous individuals produce no offspring. We also assume that mutations only occur once, at generation $n$=0.

For our generic individual, let $F_k$ be the event that $k$ of its $N$ neighbours are in state $e$. We denote by $E_{00}$, $E_{10}$, $E_{01}$, *and* $E_{11}$ the mutually exclusive events that no parents, the first parent only, the second parent only and both parents (respectively) are in state $e$. (For readability we omit the dependence on $n$ from the notation.) We will simply use $s$ for the heterozygous selective advantage ($s_{het}$).

Before we derive the difference equation for $(p_n)$, we first compute, for each specific value of $k$, the probability that an individual is heterozygous in the next generation given that $k$ of the current neighbours is heterozygous. By the law of total probability we get

$$\begin{aligned}&P(X_{n+1} = e|F_k)\\ &= 0 + P(X_{n+1} = e|E_{01}, F_k) \cdot P(E_{01}) + P(X_{n+1} = e|E_{10}, F_k) \cdot P(E_{10})\\ &\qquad\qquad + P\big(X_{n+1} = e\big|E_{11,}F_k\big) \cdot P(E_{11}).\end{aligned}$$

Using Mendelian inheritance probabilities we have $P(X_{n+1} = e|E_{10}) = \frac{1}{2}$, $P(X_{n+1} = e|E_{11}) = \frac{2}{4} = \frac{1}{2}$ and so on. Thus

$$\begin{aligned}P(X_{n+1} = e|F_k) &= \frac{1}{2} \cdot \frac{k}{N}\frac{N-k}{N-1} + \frac{1}{2} \cdot \frac{N-k}{N}\frac{k}{N-1} + \frac{2}{4} \cdot \frac{k}{N}\frac{k-1}{N-1}\\ &= \frac{1}{N(N-1)}\left(\left(N - \frac{1}{2}\right)k - \frac{1}{2}k^2\right).\end{aligned}$$

Combining these conditional probabilities with the probabilities of the conditions $F_k$ we get

$$\begin{aligned}p_{n+1} := P(X_{n+1} = e) &= \sum_{k=0}^{N} P(X_{n+1} = e|F_k) \cdot P(F_k)\\ &= \sum_{k=0}^{N} \frac{1}{N(N-1)}\left(\left(N - \frac{1}{2}\right)k - \frac{1}{2}k^2\right)\\ &\quad \cdot \binom{N}{k}\left[p_n \frac{1+s}{1+sp_n}\right]^k \left[1 - p_n \frac{1+s}{1+sp_n}\right]^{N-k}.\end{aligned}$$

Recognizing in this expression the first two moments of the binomial distribution and setting $q \coloneqq q_n \coloneqq p_n \frac{1+s}{1+sp_n}$ we get

$$p_{n+1} = \frac{N - \frac{1}{2}}{N(N-1)} N q_n - \frac{1}{2}\frac{1}{N(N-1)}(Nq_n - Nq_n^2 + N^2 q_n^2) = q_n - \frac{1}{2}q_n^2$$
$$= p_n \frac{1+s}{1+sp_n} - \frac{1}{2}\left[p_n \frac{1+s}{1+sp_n}\right]^2.$$

The equilibrium condition $p = p_{n+1} = p_n$ leads to a quadratic equation which can be solved easily.

In the case where both $s$ and $p$ are small, the product $sp$ is a second-order term, so a first-order approximation to the equilibrium value is given by

$$p = p(1+s) - \frac{1}{2}[p(1+s)]^2.$$

Therefore $p = 0$ or

$$p = \frac{2s}{(1+s)^2} \tag{A.1}$$

# ADDENDUM B SOFTWARE OPERATION INSTRUCTIONS

## B.1 INTRODUCTION

This software was created to explore the behaviour of monogenic mutations in large populations. It allows independent variation of the selection coefficients of homozygous and heterozygous individuals, as well as the population size, and the community size – that is, the local community from which an individual is likely to select a mate.

## B.2 DISCLAIMER

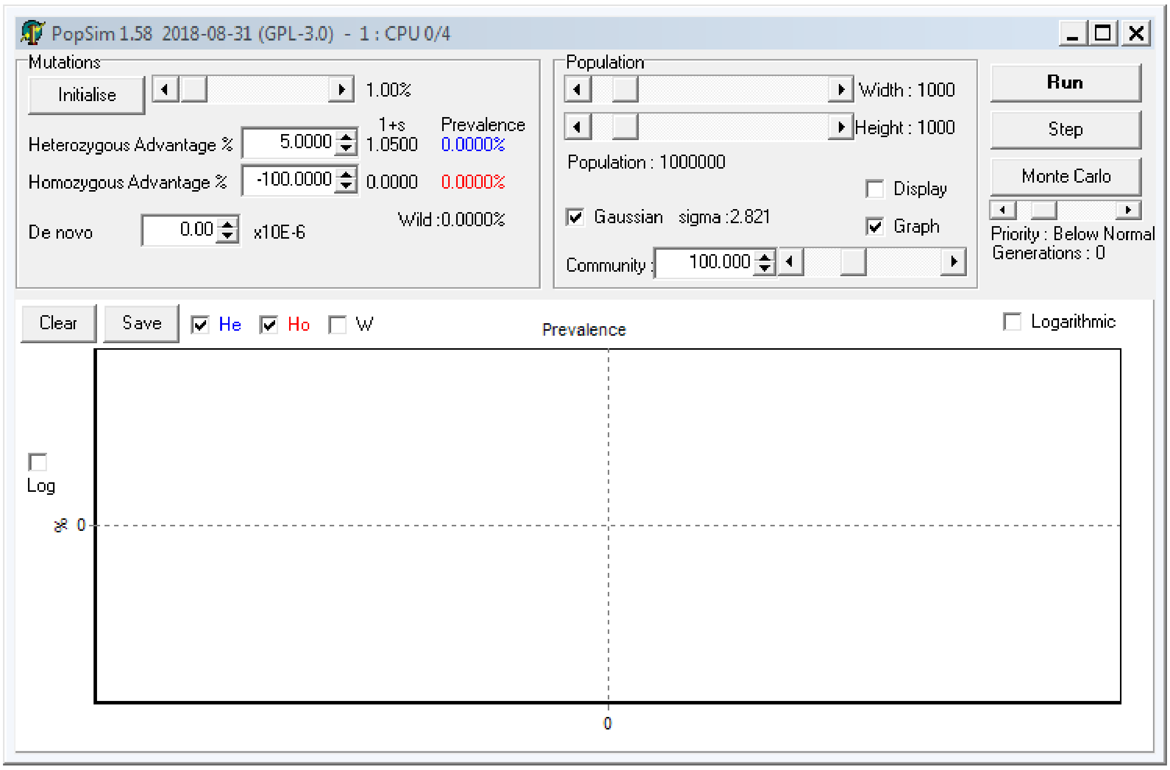
Although some effort has been devoted to make the software user-friendly and robust, it is absolutely not bullet-proof. This may be due to mere oversight, incompetence of the programmer, or possibly because of the significant weight given to speed of execution, which may occasionally result in the sacrifice of safety. It has never been the intention to reach commercial levels of polish and refinement.

## B.3 DOWNLOADING AND DEPLOYMENT

The executable is compiled for use on a PC running Windows 7 and up. It may even work on WinXP, although that has not been tested. Simply place the *.exe file in a suitable folder and run it – no installation required.

## B.4 MAIN SCREEN

The program starts by displaying a main screen looking something like this:



PopSim 1.58 2018-08-31 (GPL-3.0) - 1 : CPU 0/4

Mutations

1.00%

| | | 1+s | Prevalence |
|---|---|---|---|
| Heterozygous Advantage % | 5.0000 | 1.0500 | 0.0000% |
| Homozygous Advantage % | -100.0000 | 0.0000 | 0.0000% |

De novo 0.00 x10E-6

Wild :0.0000%

Population

Width : 1000

Height : 1000

Population : 1000000

□ Display

☑ Gaussian sigma :2.821

☑ Graph

Community : 100.000

Priority : Below Normal

Generations : 0

☑ He ☑ Ho □ W

□ Logarithmic

□ Log

**Figure B.1** – Main Screen.

On the left side, close to the top, are two text boxes containing respectively the heterozygous and homozygous selection coefficients. Note that these numbers are in percentages – to the right of the text boxes is shown the resultant selective advantage; an individual with a selective advantage of *s* will have, on average, *1+s* as many progeny as the wild type. Because it is difficult to have fewer than zero children, these numbers are constrained between -100% and +infinity (or at least an approximation thereof).

On the right, in the panel labelled ‘Population’ the size of the population can be entered – this takes the form of a two-dimensional grid of individuals, wrapping around from top to bottom and left to right to avoid any edge effects – topologically this forms a toroid. Below these can be found the community size controls: if the ‘Gaussian’ checkbox is selected, the breeding unit (local community) from which an individual is likely to select a mate takes the shape of a two-dimensional Gaussian distribution, with an effective size that of a circle with radius $2\sigma$ as defined in Equation (2.4). Proximate individuals are more likely to be selected than remote ones. If ‘Gaussian’ is unchecked, the community is a circle with size (in individuals) as shown, with all included individuals equally likely to be selected (i.e. a flat distribution).

When the ‘Display’ checkbox is checked, the population will be displayed:

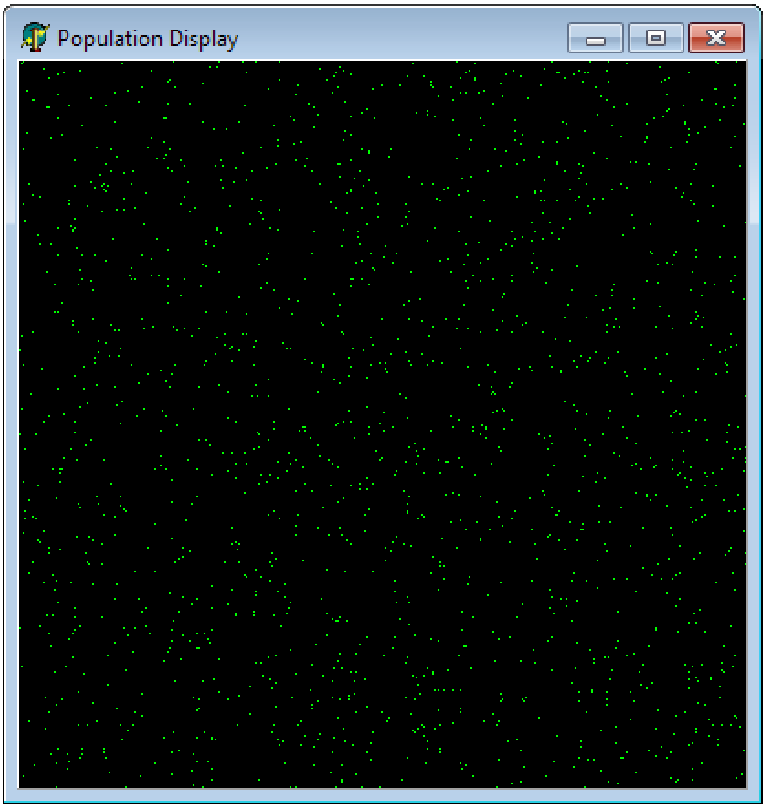


**Figure B.2** – Population Display.

In this specific case the population was initialised to 1% heterozygous prevalence by clicking on the ‘Initialise’ button at the top left, with the slider to the right of it set to 1%, of course. Heterozygous individuals are shown as green dots, homozygous ones (of which there are none in this example) as white dots, while the wild type is shown in black. Right-clicking on the population display brings up a menu which can be used to select an alternate colour scheme – black heterozygotes on a white background, with homozygotes in red – this was used to generate the images shown in Figure 4.7.

Clicking on ‘Step’ will now advance the simulation by one generation, using the selective advantages and community size settings as specified. If the ‘Graph’ checkbox, just below ‘Display’ is checked, the graph at the bottom of the main screen will also be updated with the relative prevalence numbers of the Homozygous, Heterozygous and Wild Type individuals. These graphs can be individually switched on or off by checking the applicable checkboxes at the top left of the graph.

If a non-zero number is specified in the ‘*De novo*’ box, mutations are introduced at that rate per generation (see the ‘x10E-6’ next to it, denoting *per million*). This happens randomly: in a million-element population, a *de novo* rate of $n$ x $10^{-6}$ will *on average* experience $n$ spontaneous mutations per generation, according to a Poisson probability distribution.

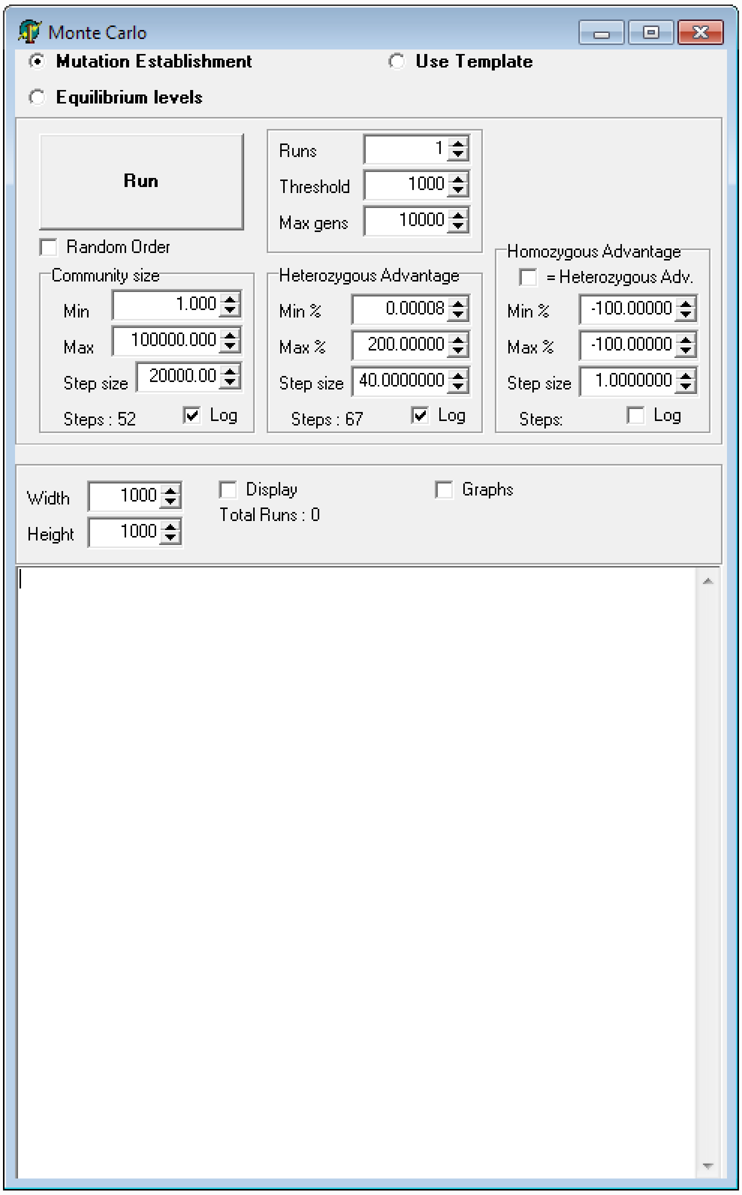
Repeatedly clicking on ‘Step’ quickly becomes tedious. Click on ‘Run’ at the top right of the main screen. This will keep stepping through generations at maximal speed until the same button is clicked again (it should be labelled ‘Stop’ while the simulation is running).

Do note that having the population display active will in general slow the simulation down significantly. It takes time to update the pixels. Not much, but there are often many of them. Also, especially when the population size exceeds that of the actual screen, the simulation may become unstable: to speed things up the screen is updated by directly addressing the video memory, which sometimes leads to unwanted effects when the display extends off-screen.

Most of the parameters can be adjusted while the simulation is running – this can be used to explore interactively.

## B.5 MONTE CARLO ANALYSIS

Because it quickly gets tiresome to keep adjusting parameters, a Monte Carlo function was created to automate this process. Click on the ‘Monte Carlo’ button to bring up the control panel:



● Mutation Establishment ○ Use Template

○ Equilibrium levels

Runs 1
Threshold 1000
Max gens 10000

□ Random Order

Community size
Min 1.000
Max 100000.000
Step size 20000.00
Steps : 52 ☑ Log

Heterozygous Advantage
Min % 0.00008
Max % 200.00000
Step size 40.0000000
Steps : 67 ☑ Log

Homozygous Advantage
□ = Heterozygous Adv.
Min % -100.00000
Max % -100.00000
Step size 1.0000000
Steps: □ Log

Width 1000
Height 1000
□ Display
Total Runs : 0
□ Graphs

**Figure B.3** – Monte Carlo screen.

The very first choice, at the top, is between ‘Mutation Establishment’ and ‘Equilibrium Levels’.

### B.5.1 Mutation establishment

When ‘Mutation Establishment’ is chosen, the screen will look similar to Figure B.3. A run involves creating a population sized as shown at the bottom left, initialised to wild type for all individuals, and then inserting a single instance of a mutation (i.e. one heterozygous carrier) into the population. After this, the simulation is run using the selected parameters until one of the stopping criteria is reached.

Multiple simulations are executed sequentially. To avert the possibility, however remote, of any given run to continue indefinitely, there has to be stopping criteria. These are: Extinction, Prevalence, and Time.

- *Extinction*: If the mutated gene becomes extinct, the simulation can be stopped, for nothing else will happen after that point.
- *Prevalence*: A mutation that does not become extinct tends to grow in prevalence. A threshold is set which, if exceeded, is considered to constitute evidence that the mutation has gained sufficient traction to make extinction unlikely, i.e. it has become established. This is set in the ‘Threshold’ text box at the top of the Monte Carlo screen.
- *Time*: It may happen that a mutation manages to linger in the population at very low prevalence levels (below the threshold set above), yet not become extinct – to prevent such situations from locking up the simulation, an upper limit on the number of generations is set in the text box labelled ‘Max gens’.

### B.5.2 Community size

The community size can be automatically stepped from the value set in the ‘Min’ box to the value in the ‘Max’ box, with increments as set in the ‘Step Size’ control below it. If the ‘Log’ checkbox is checked, the community size will be changed not linearly, but logarithmically, with the final step the size of the value in the ‘Step Size’ control. The way this is done is to start at the maximum value, then decrement by ‘Step Size’, and for each subsequent run keep decrementing by that same *ratio*, until reaching the minimum value.

### B.5.3 Heterozygous advantage

This parameter can also be changed automatically, in the same way as the community size. Understandably the logarithmic function does not work when negative values are desired.

### B.5.4 Homozygous advantage

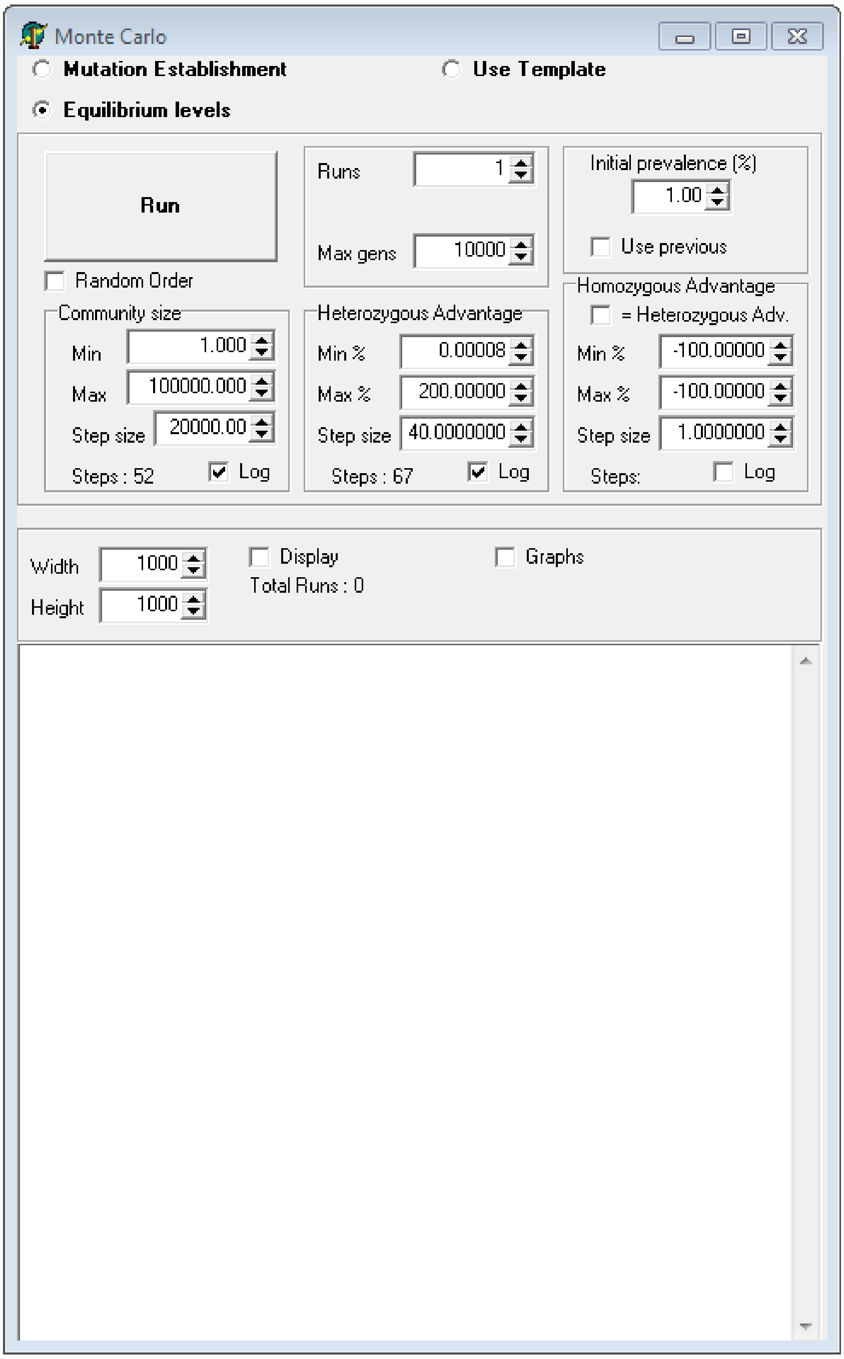
This parameter can also be automated. When the ‘= Heterozygous Adv.’ option is selected, it will be linked to the heterozygous advantage, with the same value being used.

Multiple runs with identical parameters can be executed by specifying the number of repetitions in the top middle, in the box labelled ‘Runs’.

When a run is completed, a summary of the run is displayed in the panel at the bottom of the Monte Carlo screen. This same information is also written to a text file for later analysis. The file name starts with ‘MC’, followed by the start time and date, and will be found in the same folder as the program itself.

During Monte Carlo runs it is also possible to activate the prevalence graphs, or, even worse, the population display. This is rarely a good idea if completion time is important, as it often is.

## B.6 EQUILIBRIUM LEVELS



○ Mutation Establishment
○ Use Template
● Equilibrium levels

□ Random Order

Runs 1
Max gens 10000

Initial prevalence (%) 1.00
□ Use previous

Community size
Min 1.000
Max 100000.000
Step size 20000.00
Steps : 52 ☑ Log

Heterozygous Advantage
Min % 0.00008
Max % 200.00000
Step size 40.0000000
Steps : 67 ☑ Log

Homozygous Advantage
□ = Heterozygous Adv.
Min % -100.00000
Max % -100.00000
Step size 1.0000000
Steps: □ Log

Width 1000
Height 1000
□ Display
Total Runs : 0
□ Graphs

**Figure B.4** – Equilibrium levels.

When the ‘Equilibrium levels’ option is selected, the Monte Carlo screen changes to something like Figure B.4. Most controls function similarly to the Mutation Establishment case, except that, instead of placing a single instance of a mutation in the middle of the population at the start of each cycle, the population is randomly initialized to the prevalence specified at the top right (clearly labelled ‘Initial prevalence (%)’).

The stopping criteria are Extinction, Prevalence, Time and Stability:

- *Extinction*: If the mutated gene becomes extinct, the simulation can be stopped, for nothing else will happen after that point.
- *Prevalence*: A mutation that does not become extinct tends to grow in prevalence. The simulation is stopped when 100% is reached as this is actually identical to the Extinction case above, as seem from the wild type’s perspective.
- *Time*: It may happen that a mutation manages to linger in the population at intermediate prevalence levels, yet not become extinct – to prevent such situations from locking up the simulation, an upper limit on the number of generations is set in the text box labelled ‘Max gens’. This is unlikely to happen though, because of the next item:
- *Stability*: If the code detects that the long-term prevalence levels have stabilised (somewhere between 0% and 100%, both excluded), the simulation is terminated. Due to genetic drift there is always some variability in the levels. The code for this is fairly conservative, to reduce false triggers prematurely terminating a run.

Below the ‘Initial prevalence’ box is found a checkbox named ‘Use previous’. When this is checked, the population will only be initialised by the specified prevalence when starting, or when the previous run has ended in extinction. This can be useful especially when small increments in parameter values are used, which are likely to have similar equilibrium values and spatial structures – rather than having to grow or decline from the random initial state each time, which may be quite time-consuming.

## B.7 TEMPLATES

At the start of each Monte Carlo run a template file is created, containing a summary of the run. This is to simplify repetition of an entire run. This file can be used when the ‘Use Template’ option is selected – in that case a button named ‘Template File’ appears, which is used to select the relevant template file. Do note that the option exists to also use the normal result files (starting with ‘MC’) as a template – it may just take slightly longer to analyse if it contains many repetitions of each run. Template files are simply text files – they can be edited to add or remove items, as long as care is taken to retain lines 3-5 (where the global parameters are specified).

## B.8 PROCESSOR CONTROL

Monte Carlo runs can take very long to complete, especially when the population size is large and/or many runs are desired. Some effort was expended to ensure maximal utilisation of processor resources.

During startup the program determines the effective number of processor cores in the host computer, and sets its affinity to one of these. This information is displayed in the title bar at the top of the main screen. In the example shown in Figure B.1, the program detected a four-core processor, and assigned itself to core 0 of those 4 (number 0 is actually the first core, because that is how some people, including Intel, count). If a second instance of the program is started, it will run on the next available processor core (1, in this case) etc. This is to ensure that the program does not compete with itself for processor resources, overriding the control exerted by Windows, which will happily let many processes run on one core, while others remain essentially idle. If more instances of the program are run than there are processor cores, they will of course start at zero again, and be forced to share. This is probably not a good idea.

Additionally, the priority of each process can also be controlled using the slider at the right of the main screen, just below the ‘Monte Carlo’ button. This is to enable peaceful co-existence with other programs and users, especially when all cores are being utilised by instances of this simulator. Setting the priority to ‘Below Normal’ or even ‘Idle’ means that the PC should still remain responsive, and useful for other work, even while all cores are kept 100% busy.

The above can of course also be done using the Windows Task Manager, but this way usually saves time and limits mistakes.