UNIVERSITY OF LIVERPOOL

# The Importance of Risk and Uncertainty for Humane Algorithms

Thesis submitted in accordance with the requirements of the University of Liverpool
for the degree of Doctor in Philosophy by

**Nicholas George Gray**

February 2023

# Contents

# Abstract

Computers should make our lives easier; they should complement human abilities and allow us to perform to our maximum potential. They should not be some Kafkaesque machines that require us to carefully navigate some arbitrary bureaucracy that forces us to communicate with black boxes, unsure whether we are even putting the right information into the machine or getting the right answer out. Nevertheless, the algorithms used in our daily lives are often inhumane.

One solution is for algorithms to better understand the risk and uncertainties present in the situations where they are used, within the inputs, the internal calculations and outputs. Such an approach has many potential benefits. Understanding the uncertainties can allow algorithms to make better decisions. Uncertainty in the output of an algorithm may lead to ways in which decisions can be interrogated. Allowing algorithms to deal with variability and ambiguity with their inputs means they do not need to force people into uncomfortable classifications. It is essential to compute with what we know rather than make assumptions that may be unjustified.

There are two types of uncertainty that algorithms have to deal with: aleatory uncertainty – caused by the natural variability of a system, and epistemic uncertainty – caused by a lack of knowledge. Traditionally, both types of these statistics are considered using probability theory. However, using precise probability distributions to model epistemic uncertainty can lead to illogical, inconsistent and incorrect results. This thesis considers how algorithms can compute with both types of uncertainty.

This thesis explores imprecise probabilities as a way of computing with both types of uncertainty through probability bounds analysis. The mathematics of computing with these objects is reviewed before the prospect of using an automatic uncertainty compiler to translate uncertainty naive code into code taking complete account of the uncertainty is considered. This approach can enable analysts who may be unwilling or unable to rewrite their codes to include intrusive uncertainty analysis with their algorithms.

Problems related to epistemic uncertainty within binary classifications are considered, both within the input and output of the algorithms. The problem of characterising the uncertainties associated with binary classification where there is no gold standard that can perfectly reveal the true class is explored. This problem is particularly relevant in medicine, where harm may be caused by incorrectly interpreting the result of diagnostic tests.

Epistemically uncertain inputs – in the form of intervals – within logistic regression, a popular binary classifier within statistics and machine learning, are also considered. A novel imprecise approach is shown that considers the set of possible models in an imprecise way instead of reducing the epistemic uncertainty to a single middle-of-the-road model. The approach works when there is uncertainty in the dependent and independent variables.

Finally, a chapter is devoted to considering the impact of diagnostic uncertainty within epidemiological models. Throughout the COVID-19 pandemic, governments relied on a combination of testing and epidemiological modelling to guide policy and public health interventions. However, many of the tests performed were imperfect, and it is vital to understand the impact this uncertainty has on models. This work considers that problem by creating a compartmental model with a testing and quarantining system, and can be used to analyse the effect of imperfect mass testing on the spread of the disease.

# Acknowledgements

There are many people without whom I would not have been able complete this work.

Firstly, I must thank all my colleagues at the Risk Institute for listening to me talk about my problems, supporting my research ideas, co-authoring work and providing the friendship needed to get through the past few years. Special thanks must go to Alex Wimbush, Dominic Calleja, Enrique Miralles-Dolz, Elfriede Derrer-Merk, Ander Gray, Vladimir Stepanov, Louis Clerkin, Francis Beaumont de Oliveira and Uchenna Oparaji.

To everybody associated with the EPRSC programme grant “Digital Twins for Dynamic Design”.

To Dalal Alrajeh, thank you for your support during the 6-months I spent working for Imperial College London.

To my supervisors, Scott Ferson and Marco de Angelis: The intellectual freedom you have granted me whilst pursuing this thesis has been liberating as well as a curse. Thank you for always being available to talk to me about whatever academic problem I faced.

To my parents, Andrew and Suzanne: All the support you gave me throughout my childhood, school and university journey have led to this thesis. None of it would have been possible without you.

Finally, to my wife, Harriet: Thank you for being with me through all the highs and lows whilst I have been doing this PhD. Thank you for listening to me waffle on about my work and helping me proofread, even though I am sure you didn’t know what I was talking about.

# Outline of This Thesis

Chapter 1 reviews the impact that algorithms increasing have on daily life and the ir associated risks. Several classical examples are reviewed where algorithms enabled annoyances, injustices and catastrophes. This chapter argues that better handling of risk and uncertainty would help to alleviate some of these problems and acts as the motivation for the rest of the thesis. In particular, it is crucial to compute with what we know rather than to make assumptions that may be unjustified or untenable. To this effect, this thesis aims to address twin problems problems: how uncertainty can be included directly within the calculations and how much uncertainty there is in the output.

Chapter 2 reviews probability bounds analysis (PBA) as a way of enabling algorithms to calculate their uncertainties directly. Intervals, probability boxes and confidence boxes are introduced as uncertain objects that characterise both epistemic and aleatory uncertainties.

Chater 3 introduces the concept of having a compiler that can translate source code that does not have any intrusive uncertainty analysis into code that includes PBA objects. The components that would be required to create such a compiler are considered. The advantages of such an approach and the numerous issues that would make an uncertainty compiler difficult are reviewed.

To assess the performance of binary classifiers, central to many machine learnign algorithms, one needs to be able to compare the results to the ground truth. Calculating performance statistics requires a gold standard test that determines the true classification. However, all tests are imperfect, and therefore there can be significant uncertainty associated with an imperfect gold standard. Chapter 4 explores this problem. The chapter reviews previous work addressing this problem before introducing a new method based on imprecise probabilities.

Chapter 6 takes the "no test is perfect" idea from Chapter 4 and considers how the epidemiological impact flawed testing has when testing is employed to try to stem the spread of disease. This is achieved by creating a compartmental model that can assess the dynamics of a test and quarantine system. The effect of imperfect diagnostic tests within such a system is considered for various test statistics and testing regimes.

Chapter 5 presents a novel method to include interval uncertainties within logistic regression models. This is achieved by considering the set of possible models consistent with the intervals. This approach can be used when there is uncertainty in the model's features or labels. Several algorithms are presented to estimate the set of models, and comparisons are made to methods within the literature.

Figure 1 shows all six chapters of the thesis and the connections between them.

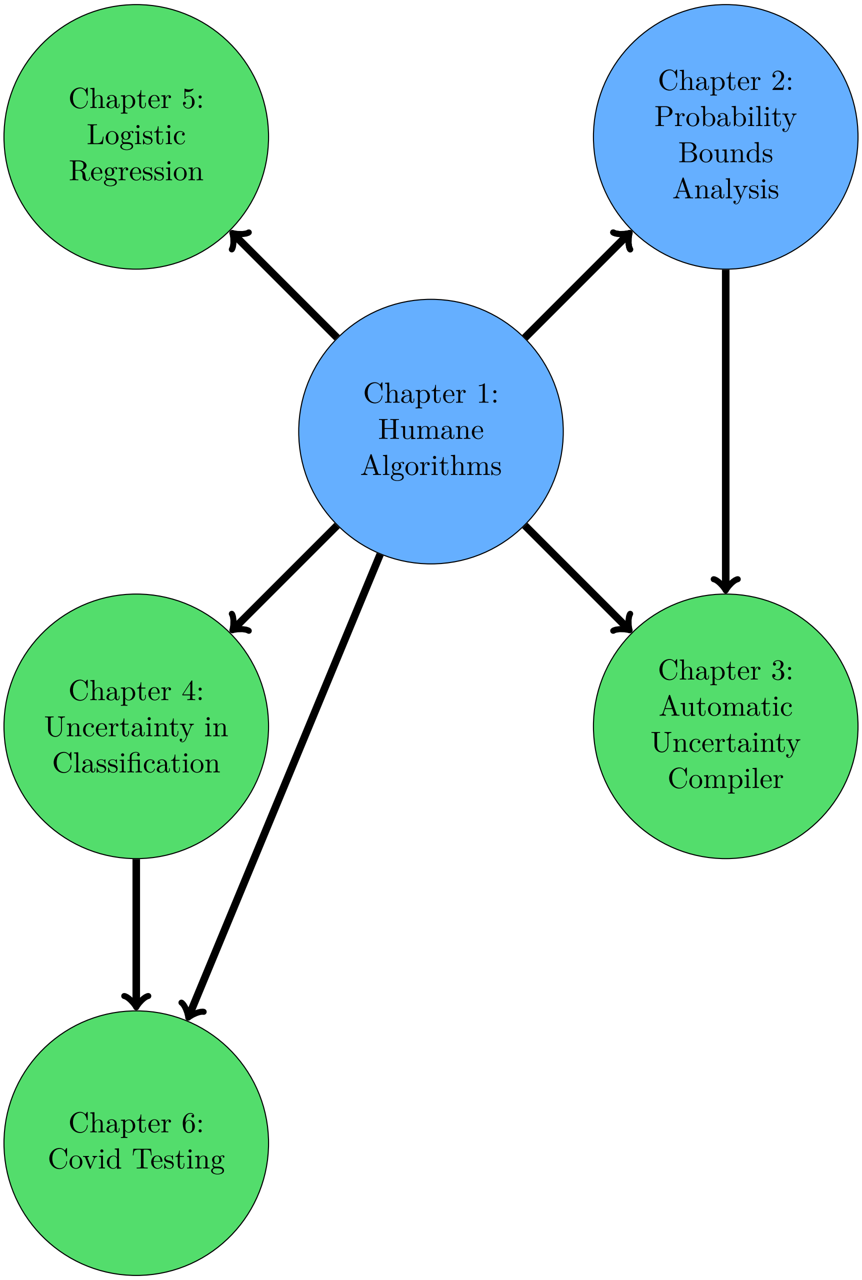


Figure 1: Connections between the chapters within this thesis. The blue chapters predominantly review the literature, whereas the green chapters represent novel contributions.

# Published Work

## Journal Articles

[1] N. Gray, D. Calleja, A. Wimbush, E. Miralles-Dolz, A. Gray, M. De Angelis, E. Derrer-Merk, B.U. Oparaji, V. Stepanov, L. Clearkin, and S. Ferson. Is “no test better than a bad test”? Impact of diagnostic uncertainty in mass testing on the spread of COVID-19. *PLoS ONE*, 15(10), 2020. doi: 10.1371/journal.pone.0240775

[2] Nicholas Gray, Scott Ferson, Marco De Angelis, Ander Gray, and Francis Baumont de Oliveira. Probability bounds analysis for Python. *Software Impacts*, 12:100246, May 2022. ISSN 26659638. doi: 10.1016/j.simpa.2022.100246

[3] Nicholas Gray, Marco De Angelis, and Scott Ferson. Towards an Automatic Uncertainty Compiler. *International Journal of Approximate Reasoning*, 2023 (*In Press*)

## Conference Papers

[4] Nick Gray, Marco De Angelis, Dominic Calleja, and Scott Ferson. A Problem in the Bayesian Analysis of Data without Gold Standards. In *29th European Safety and Reliability Conference*, pages 2628–2634, Hanover, Germany, 2019

[5] Nick Gray, Marco De Angelis, and Scott Ferson. Computing With Uncertainty: Introducing Puffin the Automatic Uncertainty Compiler. In V. Papadopoulos M. Papadrakakis, G. Stefanou, editor, *3rd ECCOMAS Thematic Conference on Uncertainty Quantification in Computational Sciences and Engineering*, pages 487–497, Heraklion, Greece, 2019. doi: 10.7712/120219.6354.18702

## Open Source Software

[6] Probability Bounds Analysis for Python. https://pypi.org/project/pba/{or}https://github.com/Institute-for-Risk-and-Uncertainty/pba-for-python/,

## Named Author

[7] Alexander Wimbush, Nicholas Gray, and Scott Ferson. Singhing with confidence: Visualising the performance of confidence procedures. *Journal of Statistical Computation and Simulation*, pages 1–17, March 2022. ISSN 0094-9655, 1563-5163. doi: 10.1080/00949655.2022.2044814

[8] Francis J. Baumont de Oliveira, Scott Ferson, Ronald A. D. Dyer, Jens M. H. Thomas, Paul D. Myers, and Nicholas G. Gray. How High Is High Enough? Assessing Financial Risk for Vertical Farms Using Imprecise Probability. *Sustainability*, 14(9):5676, May 2022. ISSN 2071-1050. doi: 10.3390/su14095676

## Pre-prints

[9] Nicholas Gray, Marco De Angelis, and Scott Ferson. The Creation of Puffin, the Automatic Uncertainty Compiler. *arXiv:2110.10153 [cs, stat]*, October 2021. http://arxiv.org/abs/2110.10153

[10] Nicholas Gray and Scott Ferson. Logistic Regression Through the Veil of Imprecise Data. *arXiv:2106.00492 [stat]*, June 2021. http://arxiv.org/abs/2106.00492

# A Note on the Word "Algorithm"

The Oxford English Dictionary defines the world algorithm (in a computer science and mathematics context) to mean

> *A procedure or set of rules used in calculation and problem-solving; a precisely defined set of mathematical or logical operations for the performance of a particular task.*

In a decision making context algorthims tend to be either:

- *rule-based algorithms*: human-defined instructions intended to be direct and unambiguous, or
- *machine learning algorithms*: where the algorithm comes up with its own instructions in order to solve the task.

These algorithms often perform one of a few basic tasks:

- prioritisation – finding a ranking over a given number of objects to return the best,
- classification – picking a catagory,
- association – finding links between objects, and
- filtering – isolating the important information [11].

An alternative approach defines algorithm singularly and anthropomorphically to describe the collection of different rule-based or machine learning systems that form a black box system that takes an input, performs some mathematical and computer science wizardry and returns an output [12]. The YouTube algorithm decides on what video to recommend next. Banks use algorithms to decide whether or not to give someone a credit card. Within a self-driving car it is the algorithm that is doing the driving, despite the fact that there will be numerous individual algorithms that decide on the speed, the direction and monitor the traffic conditions.

It is within this anthropomorphic definition that algorithms need to be made more humane.

# Chapter 1

# Humane Algorithms

In Issac Asimov's seminal science fiction book *I, Robot*, the *Handbook of Robotics (56th Edition, 2058AD)* states that the first law of robotics is: "A robot may not injure a human being or, through inaction, allow a human being to come to harm." [13] Although some suggest that Asimov's Laws may be a simplistic view of machine ethics [14], the first law is a pertinent starting point when it comes to discussing the risk posed by algorithms. Increasingly inhumane algorithms adjuidcate more and more of our daily lives, with potential harms ranging from irritation to injustice, from confustion to catastrophe.

There is even evidence that people are beginning to move from "algorithm aversion", preferring advice from a human over advice from an algorithm, to "algorithm appreciation" and going out of their way to do things using automated systems [15]. This trend does depend on the context of the decision, with human decision making still preferred in medicine, economics and education [16–18].

Then there is the danger of humans treating algorithms as epistemic superiors who can perform tasks far better than other humans and trusting in the algorithm's advice over their own decision making. Take, for instance, the numerous examples of people who have driven cars into perilous, and sometimes fatal, situations because their satellite navigation system instructed them to do so [19, p. 15].

There is a need to ensure that algorithms are humane and work and play well with humans, anticipating and serving their needs and frailties. There are several fundamental issues that a humane algorithm would need to address. The review Jobin et al. [20] revealed that there is an emerging consensus amoung the many principles and guidelines published for ethical artificial intelligence. The five ethical issues they highlight are: transparency (including explainability), justice and fairness, non-maleficence, responsibility and accountability, and privacy. Although, by definition, no humane algorithm would be maleficent. Humane algorithms do not just include the issues related to AI ethics. However, they must also consider the numbers used within their calculations and the context for the decisions as essential issues. This chapter reviews such issues and highlights that careful consideration of risk and uncertainty is fundamental in creating humane algorithms.

## 1.1 The Numbers of the Future

“Prey, Mr. Babbage, if you put into the machine the wrong numbers will the right answers come out”. This question was put to Charles Babbage twice, by members of both houses of the UK Parliament, upon his invention of the difference engine [21], which can be considered as a very early ancestor of today’s computers. Babbage was “not able rightly to apprehend the kind of confusion of ideas that could provoke such a question.” Yet, similar problems still occur even though almost two centuries have passed between Babbage’s conversations with the parliamentarians. It is still the case that if the input is incorrect, the algorithm will still produce an output that will be wrong, but with the added air of algorithmic authority, often referred to as “garbage in, garbage out”.

What Baggage could not have possibly imagined is that the distant descendant of the machines that he created would be so ubiquotous within today’s society. He would not be able rightly to apprehend that people are still putting the wrong numbers into the machine and expecting the right answers to come out.

Modern science and engineering is all about numerical calculation. With the inexorable growth of computer power, more of these calculations are being undertaken with more complex computer simulations. These developments mean that new computation-intensive and machine learning technologies are being explored which promise to revolutionise engineering by replacing the magic of engineers with algorithmic witchcraft.

The numbers used within scientific calculations do not represent so abstract philosophical or numerological constructs but instead are counts, measurements or other representations of the real world. For instance, if an engineer needed to use the number 5 they do not mean some vague concept of *fiveness* but instead a real value likely to have some associated unit and possibly some uncertainty. However, the computer languages used to make these calculations have very few good options to take units and uncertainties into account. It would make no sense to add 5 meters and 4 seconds, nor would it be possible to calculate the sine of something which is not an angle, yet computers blind to units would allow such calculations. It is left to the programmer to understand the units within their calculations and ensure that there are no errors.

Several high profile disasters and embarrassments have been due to algorithms being ignorant of units. In 1983, Air Canada Flight 143 ran out of fuel mid-flight because of conversion errors between pounds, kilograms and litres [22]. In 1999, NASA lost contact with the Mars Climate Orbiter during an orbital insertion manoeuvre after its journey from Earth to Mars when a sensor reported data in imperial units (pound-seconds) that the algorithm wanted in metric units (newton-seconds) [23]. In a pharmaceutical context, such unit errors can have deadly consequences. In the United States, between 3,000 and 4,000 children need emergency hospital treatment each year due to mistakes made by caregivers administering drugs, some of which can be attributed to conversion errors between unit systems [24].

Such unit conversion errors are often labelled as human errors, yet the human is only one part of the system. The Mars Climate Orbiter mishap report recommended that the project engineers "verify the consistent use of units throughout the spacecraft design and operations" [23, p. 7], again emphasising the human's role in preventing future incidents. However, the computers performing the calculations could have been aware of the units. Instead of requiring the pilots of Flight 143 to calculate the fuel load in a unit system that they were not familiar with, the onboard computer could have been more humane and accepted the fuel load in either unit system.

Equally, the algorithms that we currently use often do not account for uncertainty in natural ways. There are two types of uncertainty:

- *Aleatory uncertainty* caused by natural variability of a system, such as the number of a dice after it has been rolled, and
- *Epistemic uncertainty* caused by a lack of knowledge.

As will be discussed within Chapter 2, aleatory uncertainty can be naturally characterised by probability distributions, whereas intervals naturally characterise epistemic uncertainty. These uncertainties may arise from many sources; epistemic uncertainty might arise from measurement errors, missign data or censoring of data. Linguistics contains many natural ways of characterising uncertainty through statements such as 'about 5' or 'almost 3'. Mathematical contructions allow us to formally define these uncertainties, for instance within machine learning aleatory uncertainty is usually presented using probability estimates. Epistemic uncertainty is often imagined to be covered using Bayesian methods; however, this is debatable at best because it is challenging to represent the lack of information using a single probability distribution [25].

A humane algorithm should be able to handle numbers in many intelligible formats, accepting quantities in any appropriate units and checking that their inputs have the correct dimensions. For example, an input field expecting a distance should be able to correctly interpret inputs giving kilometres, miles, feet or metres without forcing a person to make the conversion. An entry without units or non-conforming units such as grams should precipitate follow-up questions. Input fields expecting numerical quantities should be able to accept and process expressions of uncertainty in the inputs, such as ranges or plus-or-minus statements. Interfaces should accept inputs, including linguistic hedges such as ‘about’, ‘no more than’, or ‘up to’ to modify numerical quantities [26]. Uncertainty in, uncertainty out is preferable to garbage in, garbage out.

## 1.2 Risk Awareness

Algorithms have no idea of the significance of the calculations that they are performing. Like Babbage’s difference engine, algorithms will take any input and thoughtlessly produce some output. This can have severe consequences for algorithms that make decisions in high-risk contexts.

The Therac-25 was a software-controlled radiation therapy machine produced by Atomic Energy of Canada Limited (AECL) in the early 1980s*. Between 1985 and 1987, at four medical centres, at least six patients received potentially lethal doses of beta radiation. The Therac machine had two models that could be used to treat patients. The first was an electron therapy mode in which a low current beam of electrons at various energies (5 MeV to 25 MeV) is directed at the treatment area. The machine could also be used for photon therapy in which a high-current high-energy (25 MeV) electron beam collides with a target, and the emitted X-rays are targetted at the treatment area. On six different occasions, the high-current high-energy electron beam that should be used to produce X-rays was instead targeted directly at the patients, delivering 100 times the intended dose of radiation, causing radiation burns, radiation poisoning and, in three cases, death.

Several software engineering failures were directly responsible for the accidents, one of which was that the ‘Set-Up Test’ procedure, which was supposed to check whether the machine was positioned and programmed safely, used an 8-bit variable to store the safety value of the machine. If the value was 0, then the machine would be ready for use. Any other value the programme would return an error. Every time the test was performed on a not-safe machine, the value would be incremented by 1. If the procedure was performed 256 times, the safety value would overflow and the counter would return 0, meaning that it would appear to the algorithm that all the safety checks had been passed when they had not been. Other software errors caused the machine to ignore changes or corrections made by the operator. Interestingly, the investigation into the failures found that “The error was most

*The factual information for this discussion comes from [27–30].

likely to occur if the operator was experienced and quick at editing the input" [27, p. 151].

Whilst an algorithmic aeon might have passed since the creation of Therac-25, there are many important lessons from the incidents that are still pertinent today. The error messages that appeared on the interface ("Malfunction 54" or "H-tilt") were simple, obscure and did not imply the potential risk the error was preventing. The user interface was confusing. The manufacturers were overconfident in the machine's performance and refused to consider the possibility that the algorithm made a mistake. They even used a probabilistic risk assessment to justify its safety.

As with the unit errors, incidents like those produced by the Therac-25 machine are often attributed to human factors, assuming that the machine's operators should be omniscient about the operation of the machine and that they can understand the error messages and respond accordingly. Most modern programming languages allow programmers to create better error protection and messages. It is relatively simple to create error messages intelligible to the software's end-users. Such processes should not be an afterthought; they should be designed into the software itself. This is particularly important when decisions may carry significant risk.

We see the same problem when it comes to dispensing medicines, when a nurse needs to work out how much drug a patient can safely consume, it is often a trivial calculation. It is also trivial for the nurse to have made a mistake in the calculation. Simply misplacing a decimal point could result in an ineffective dosage or one that may be fatal. When such incidents have the most dire consequences, analysis of the causes often overlook the role that the calculator played within the incident, focusing on clinical issues and, again, human factors [31]. The algorithms that perform the calculation are considered unlikely to have made a mistake – ignoring the possibility of round-off, overflow, division-by-zero, or similar errors impacting the calculation. Fortunately, it is possible to design calculators to deal with some of these issues. Calculators can be designed that can detect and block such simple number entry errors [32]. Context-aware algorithms can "sense check" results and help to prevent harm being done as a result of incorrect calculations.

Within more complicated systems, algorithms need to be risk aware and fail in ways that do not make the situation more dangerous. For example, Air France Flight 447 (AF447) was an Airbus A330 aircraft flying between Rio De Janeiro, Brazil and Paris, France, on 1st June 2009. There were no prior mechanical faults on the aircraft, nor was the weather unusual for the time of year, although there was some mild turbulence experienced. Midway across the Atlantic, the aircraft experienced icing conditions, and ice began to build up inside the pitot tubes that measure how fast the aircraft is moving. Faced with uncertain speed information, the autopilot disengaged, passing over control of the aircraft to the two pilots in the cockpit. Unused to flying the aircraft at 38,000 ft, the pilots in control of the aircraft made a series of errors that eventually put the aircraft into a stall and even with the stall warning sounding, the pilots were unable to recognise their error until it was too late and the aircraft crashed into the ocean killing all 228 people on board the aircraft [33].

The algorithm's approach to dealing with the uncertain speed information was to give up, leaving all decisions to the pilots to troubleshoot. The autopilot could have helped the humans. Whilst the pilots did correctly identify the lack of airspeed, neither of the two copilots called the "Unreliable Indicated Airspeed" procedure [33, p. 198]. The autopilot could have helped them understand the situation by guiding them to the correct checklist or by helping them understand the steps they needed to take to fly safely. Giving the pilots contextualised information such as "Disconnect the Autothrust, Autopilot, and Flight Director"[†] or "Keep the aircraft away from the low-speed and high-speed ends of the flight envelope"[‡] could have provided the pilots with helpful advice to keep the aircraft from crashing.

Safe human-computer interaction will be a critical component of any humane algorithm.

## 1.3 Justice and Fairness

Humane algorithms will need to be both fair and just. Whilst providing precise definitions for these terms can be challenging, it is obvious when algorithms are making unfair decisions and helping to entrench injustice. There are numerous different ways in which algorithms can propagate injustices.

Digital discrimination is a form of injustice in which algorithmic decisions treat users unfairly, unethically, or just differently based upon protected characteristics, including[§]: age, disability and sexual orientation [35]. This discrimination may be: direct and explicit, such as having a hiring policy that excludes young women with children; direct but implicit, such as having a hiring policy that excludes people who have recently taken a career break – disproportionately likely to be young women with children; or indirect, having a neutral hiring policy that still makes it difficult for young women to be employed [36]. If the algorithm in question is a black-box machine learning system such that no human is able to see

[†] Whilst the autopilot and autothrust systems on the aircraft automatically disconnected the Flight Director computer system stayed active. The report into the incident hypothesises that "Flight Director indications [may have] led the crew to believe that their actions were appropriate, even though they were not" [33, p. 200].

[‡] This advice is taken from SKYbrary [34].

[§] See https://www.equalityhumanrights.com/en/equality-act/protected-characteristics for complete list the created rules, it is likely to be the case that it is only possible to detect the bias after the algorithm has been let loose in the real world.

Numerous examples of algorithms that have been accused of sexism: Apple and Goldman Sachs were accused of systematically granting men higher credit card limits than women [37]. Female content creators have reportedly pretended to be male on Instagram to get around community guidelines that "disproportionately affect women" [38, 39]. Music streaming services are disproportionately likely to recommend male artists over female artists [40]. Many algorithms indirectly discriminate against women by working on a "one-size-fits-men" approach making it difficult for women to interact with them [41, Chaper 8].

One-size-fits-men could easily be one-size-fits-*white*-men as there are also numerous examples of algorithmic racism: Sweeney [42] found that Google displays adverts such as "Have you ever been arrested?" when searching for black-sounding names. Twitter has had to apologise for a facial recognition algorithm that automatically cropped out black faces [43]. Perhaps the most notorious example of a racist algorithm is COMPAS.

COMPAS was an algorithm used within the justice system of several US states that tried to predict the risk of recidivism; when studied, it was found to have a racial bias problem [44, 45]. It was more likely to score black defendants as higher risk than white defendants, even when the white defendant had a more egregious criminal history. While the overall accuracy for both black and white defendants was similar, the algorithm misclassified more black defendants as high-risk than white defendants and, conversely, misclassified more white defendants as low-risk than black defendants.

The COMPAS's racial bias problem occurred even though race was not included as a predictive feature [46]. The algorithm predictions were based upon a questionnaire with 137 questions, including: "Was one of your parents ... ever sent to jail or prison?", "Is there much crime in your neighborhood?" and "How many of your friends/acquaintances are taking illegal drugs regularly?". Such questions have inherent racial bias problems, are backed up by years of institutional police racism, and are implicitly discriminatory [47, 48]. Each of these questions is likely to be a proxy for race, meaning race factors are likely to be included within the algorithm implicitly.

There are numerous challenges when it comes to trying to assess the risk of bias, from the fact that it is often difficult to define what exactly a fair algorithm would be and, therefore, there are many different ways in which fairness can be assessed [49, 50]. COMPAS passes some fairness tests such as predictive parity and is equally accurate for both white and black groups [51]. Whether that accuracy should be considered good or not is a separate ethical question. Dressel and Farid [52] found that COMPAS performs similarly to recidivism predictions made by humans with little or no criminal justice experience. Their study also found that the humans were also just as likely to fall into the same racial bias as the algorithm, yielding more false positives for black defendants than white defendants. It is also not the case that every judicial judgement made by humans can be considered fair – judges are prone to human error, have a long history of racial bias [53] and are more much likely to give favourable decisions after lunch than before they have eaten [54]. Such need for algorithmic fairness may be the result of the fact that people are far less tolerant of mistakes made by algorithms than similar mistakes made by humans [55].

One of the problems with algorithms such as COMPAS is that there is no gold standard or ground truth that can be used to verify the true classification. For COMPAS, in theory, it is possible to know whether a high-risk prisoner did go on to reoffend, but, as authorities are unwilling to release defendants on parole if the algorithm determines that they are high risk, it would be impossible to know whether the algorithm was accurate or not. If a defendant fails the test and thus remains imprisoned, there is no available data on whether they would have reoffended had they been released. Similarly, credit card companies that use algorithms that determine the creditworthiness of applicants do not like giving cards to individuals for whom there is a perceived risk of default. Therefore it may be impossible to say whether these cases were predicted accurately.

Algorithms that try to predict subjective data may also suffer from the fact that there is often no gold standard. For example, a company might use an algorithm to filter applicants for a job application. Such an algorithm may output a list of “good candidates” for the role, but there is unlikely to be a precise definition of what a good candidate is. What one recruiter considers a good candidate may be different to another recruiter’s definition; how it is measured is likely to change over time and, as touched upon above, be subject to either direct or indirect biases.

In both cases, the lack of a gold standard means that any performance statistic (accuracy, true positive rate, false positive rate, etc.) used to justify the algorithm’s fairness is likely to have some hidden uncertainty. Correct quantification of this uncertainty may lead to insights into the fairness of the algorithm. For example, using such insights may reveal that there is significant uncertainty in COMPAS’s accuracy predictions which could impact the calculations made that Flores et al. [51] and Angwin et al. [44] use to debate the fairness of the algorithm.

A related problem is that of feedback loops. The PredPol algorithm tries to predict areas in which high rates of crime will occur meaning that limited police resources can be appropriately used [56]. Lum and Isaac [56] note that in areas where PredPol recommends that officers spend time on the beat, there are more likely to be arrests, meaning that such areas are more likely to be recommended by PredPol than other areas. This feedback loop leads to disproportionate policing of historically over-policed communities – often impoverished and with higher black populations than areas that were not targetted – which leads to harm as individuals living in these areas may feel oppressed. As with COMPAS, these injustices are mirrored when a human, not an algorithm, makes the decision. The reduction of epistemic uncertainty in areas recommended by PredPol is treated as thought it is variability and the epistemic uncertainty in non-recommended areas is ignored.

Another sort of injustice is what Goodman and Flaxman [57] define as 'uncertainty bias', a problem occurring when the following two conditions are met:

- One group is underrepresented in the sample, so there is more uncertainty associated with predictions about that group.
- The algorithm is risk-averse, so it will – all things being equal – prefer to make positive decisions based upon predictions about which it is more confident or have less uncertainty and default to a negative decision otherwise.

Goodman and Flaxman then give the example of an algorithm deciding whether to give an individual a loan. If in a population the number of white people is 95%, then there will inevitably be a smaller sample size of non-whites for the algorithm to learn about, then, due to the smaller sample size, the uncertainty about any prediction within that group will be higher. This would be the case regardless of any systematic biases within the data set, which could add another layer of injustice. This bias is artificial; it results from treating the epistemic uncertainty caused by the lack of information about the group as though it was aleatory uncertainty.

Such a trends are similar to the False Confidence Theorem defined by Balch et al. [58]: that, when modelling large epistemic uncertainty using probabilities, there is a risk that a false hypothesises receives arbitrarily large posterior probability. i.e. as the epistemic uncertainty about a prediction increases, the probability of the event occurring appear to decrease to zero. This effect is counterintuitive; as the uncertainty increases, the decision should become more and more unsure, not the other way around. The false confidence theorem can be remedied by using imprecise probabilities to model both the epistemic and aleatory uncertainty with the calculations, as will be discussed in Section 2.6.2. Another solution to this problem is that instead of forcing an algorithm to make a decision when faced with significant epistemic uncertainty, it could instead simply return "I don't know", although this would have to occur in a way which protects against the risk awareness issues discussed in Section 1.2.

Another injustice that can occur when people are forced to interact with inaccesible algorithms. This is an increasingly pertinant issue as participation in today's society depends on access to and facility with computers and the internet. We are effectively forcing people to engage with technology in ways that many find difficult. Such trends risk excluding those who struggle to interact with algorithms, creating a phenomenon known as the "Digital Divide" [59], and its severity appears to be worsening. The general public may not even be aware of the algorithms. For example Gran et al. [60] found that 61% of the Norwegian population has little to no awareness of the algorithms that increasingly mediate their participation in public life.

Essential interactions in banking, health care, government agencies, and even restaurants and social gatherings are mediated through automated attendants, chatbots, robocalls, and social media. Considerable effort has been invested in making these interactions accessible to everyone through various assistive technologies. Nevertheless, because circumstances and public policies constantly evolve, the interfaces are often in flux, and it has proven to be challenging to keep websites current and fully functional. This problem is severe for the government as more and more public services are mediated over the internet.

Inputs are sometimes unnecessarily identified as required (or only needed so that the data handler can sell it to advertisers). Declining to answer by leaving a field blank, perhaps because the correct answer is unknown, should not necessarily prevent a well-written algorithm from using the available information. Similarly, multiple-choice input fields should accept multiple, none-of-the-above, mixed, and uncertain information. For instance, binary gender assignment can be unnecessarily troubling for transgender and intersex individuals. Race information is also often improperly characterised as white/non-white or white/black/other.

For example, Osler et al. [61] present a logistic regression model to assess the risk of death following a burn injury. The data they used within their model include binary labels for both race (0 for "non-whites" and 1 for "whites") and gender (females are 0 and males are 1). Such calculations are clearly uncomfortable, but are not unique to Osler et al. All of the example datasets in the logistic regression textbook by Hosmer Jr et al. [62, pp. 22–33] (which does reference the Olser dataset) contain similar characterisations. Such classifications significantly obscure the diversity of humanity. It is clearly the case that human diversity does not neatly categorise into white, black or other.

Requiring inputs to have a particular format, or to be expressed on a particular scale or in particular measurement units, is unnecessary and inappropriate in many situations. Such tyranny of input boxes may lead to *hermeneutic injustice*: where a person is disadvantaged because they are unable to make their experiences intelligible [63, 64]. For example, a person may not communicate their symptoms accurately to a medical decision support algorithm. The algorithm might want to know how long the patient has had a cough but only accepts precise answers, whereas the patient might only provide an estimate such as “at least three weeks but no more than two months”. The algorithm might ask questions in medicalese that the patient might not understand, such as “do you have arthralgia?”. Alternatively, the patient may simply not know the answer to the question.

There will always be some form of injustice; life cannot always be fair to everyone, and algorithms will perpetrate some injustices. This is especially true when the use of such algorithms gives an aura of authority to the harmful decisions. However, entrenching pre-existing injustices only makes the situation worse. Correctly identifying the risk of injustice for a particular group would be beneficial. As would improving the accessibility of algorithms by enabling natural inputs to be used and stopping forced categorisations. Humane algorithms should not be tyrannical.

## 1.4 Transparency

When it comes to transparency there are two topics that are important to discuss: explainability and provenance.

The European Union’s General Data Protection Regulation [65] aims to give individuals control of their personal data and how it is processed by corporations. Articles 15(1)h and 22(1,4) place restrictions on how algorithmic decisions can be made. They give individuals the right to “obtain meaningful information about the logic involved in a decision” made by an algorithm and the right to object to decisions being made solely by an algorithm. The field of explainable artificial intelligence (XAI) has arrisen in order to try to solve this problem [57, 66, 67]. Some argue that explainability is not the remedy that is needed to protect against malicious algorithms [68]. At the same time, explainability may not be possible, and there are situations where a complete understanding of the algorithm is neither desired nor required [64].

Provenance records the origin of an algorithm or dataset, what assumptions have been made, if there have been any alterations, and who owns and administors the algorithm/-data. Access to provenance will be an important characteristic for humane algorithms as it will enable increased transparency about the algorithms’ decision-making process. Knowing the provenance can also have other benefits. For instance, the provenance of information on social networks can help to protect against the spread of misinformation on social media websites [69]. There is more disinformation about COVID-19 than information on social media [70] and much of this can be sourced to spurious claims made by only twelve individuals [71, 72]. Similarly, 69% of misinformation about climate change can be traced back to ten sources [73]. Being aware of the provenance of the claims may allow algorithms to be created that are able to detect dangerous falsehoods when recommending content. Of course, this depends on social media companies caring more about user welfare than the shareholder value generated by ad revenue that online spaces colonised by hate and misinformation make.

Due to the rise of deepfakes, among other technologies, a video of someone confessing to a crime may not be valid evidence unless the provenance of the video is known. The lawyers of the future may argue over where the video was created, how it has been edited, where has it been stored, who has had access to, it in order to validate its credibility. Alibis may be used to exonerate defendants by arguing that they could not have made the video.

In medicine, machine learning algorithms are becoming increasingly important tools to aid diagnosis. In future regulators may ask: why a particular machine learning algorithm has been used? Where did the training data come from? What was the data cleaning process? Who made what assumptions and why? Assuring the safety of such algorithms may require knowledge of the provenance.

Another area where transparancy will be critical is when it comes to the increasingly complicated models that are being developed within various application areas. In engineering there has been much hype about digital twins which are supposed to be a virtual representation of a real world object (see [74, Sec. 2.2.3.3] or [75]). Various definitions exist of digital twins [76, 77] but the ultimate aim is for the digital twin and its physical twin to be indistinguishable. However, it is worth considering that most humans twins are not identical and nor will digital twins be [78].

A digital twin will be a collection of different algorithms and other components that are bounded up into a single system. Each of these individual models will have their own provenance that will need to be curated. Additionally, each individual model will have some associated assumptions that will be key to the operation of the algorithm. Transparency will be critical in explaining why these decisions were made in the creation of the model and a where such decisions have an impact on the identicality of the twin. It will be important to ask the question, “how can I trust that this really is a twin?”. This will be particularly important for digital twins of safety critical systems or those that represent difficult to access objects. Being transparent about the provenance of the twin, what assumptions went into the model and being able to explain why the twin is acting as it is will be cruitial.

Several assumptions fundamental to machine learning are often made without justification. For example, assuming data is *independant and identically distributed* (i.i.d.) is a critical inbuilt assumption for many machine learning algorithms. The i.i.d. assumption is that all instances come from the same underlying probability distribution, and all instances are mutually independent. The i.i.d. assumption is often violated in practice. Equally, missing data is often assumed to be *missing at random* or *missing completely at random*, even when these justifications may not be the case.

The machine learning community implicitly assumes that such violations do not impact the performance of algorithms or that the violations are only slight and therefore do not impact the final analysis [79]. Taking the view that if an algorithm is well-performing, then adhering to precise statistical definitions is not necessary. This may be suitable for many cases, but analysts should be transparent about what assumptions they are making when creating their algorithms and these assumptions shoukd be recorded as part of the provenance of the algorithm.

Increasingly algorithms are being used to aid diagnosis. For example, image recognition algorithms can be used to detect breast cancer [80–82]. No diagnostic test is perfect, including those performed by algorithms, so they will give false positives and negatives. The performance of the tests is often expressed using the following statistics: sensitivity ($s$), also known as recall – the proportion of sick individuals correctly given a positive result; specificity ($t$) – the proportion of well individuals given a negative result; and, positive predictive value ($\psi$) – the likelihood of being a true positive after receiving a positive result; accuracy ($a$) – the total number of individuals who received the correct result. In the machine learning literature, sensitivity is often known as recall, and $PPV$ is often known as positive predictive value.

Unfortunately, the result of a test, along with these statistics, may not give an accurate representation of the performance of the test as the prevalence ($p$) of the disease matters. For example, the positive predictive value is dependent on the prevalence of the disease through Bayes rule,

$$\psi = \frac{ps}{ps + (1-p)(1-t)}.$$

Bayes rule means that how likely it is a positive is a true positive is dependent on how likely it is that the subject was positive in the first place. As such, in situations where there are not many sick individuals to begin with, it is likely to be the case that the majority of the positive results from the test will be false positives. Being transparent about the presumed prevalence is important for explaining why the algorithm came up with a particular decision.

AS it is often preferred that diagnostic algorithms are risk-averse, meaning that analysts often design algorithms such that the number of false negatives is minimised even though it results in more healthy patients being given a false positive. For instance, a medical image recognition algorithm used to detect breast cancer as part of a screening programme is likely to be optimised so that as few patients with cancer are given the all-clear, even though such an approach means more patients get false-positive results. Since false positives are not harmless occurrences, causing psychological distress and the risk of unnecessary medical interventions, the fact that the $PPV$ of the algorithm is low needs to be communicated to patients.

It has been observed that medical practitioners struggle to communicate this information to patients [83]. Improvements can be made by presenting information in better ways for people to understand. Natural frequencies are a more intuitive way for people to understand the Bayesian reasoning required to calculate the probability that someone has a disease given they have received a positive result [84, 85]. This can be done by using statements such as inputting the prevalence as "2 out of 1000 women have breast cancer" instead of $p = 0.002$ or "9 out of 98 women who test positive will have breast cancer" instead of $PPV = 0.092$. Icon arrays are another approach that can be used to express information from algorithms [86]. They show natural frequency information using pictograms that are simple for people to understand even if they have poor numeracy skills [87].

More verbose statements could be used in addition to these giving more information to the algorithm's users, such as the assumption made during the calculation. For example, instead of a recidivism algorithm outputting "high risk," it could instead output a statement such as:

*Out of 100 defendants with this test score, between 60 and 70 will reoffend within two years, this calculation has been made assuming that around 3 out of 10 defendants get rearrested within this period.*

Such a statement expresses the likelihood of the defendant reoffending, including the associated uncertainty about the prediction and telling the user what the assumed prevalence of reoffending was. This output would allow an avenue for appeal since the assumed base rate used within the prediction could be challenged.

Clearly expressing the uncertainty may also help to ensure algorithmic accountability. Imagine that an algorithm used to test whether someone should get a loan gave the following statement:

*Between 5 and 95 out of 100 similar applicants who get a loan will default on their payments.*

In this case, we can see that there is considerable uncertainty about the prediction and as a result, it would be difficult to make a definitive prediction about how likely it is that the person will default. As such, the person who makes the final decision would know the uncertainty about the algorithm and be able to adjust their decision accordingly. Verbose statements need to be carefully constructed, ensuring that the gist of the output is clear to the users and not obstructed with unnecessary information. Expressing the uncertainty within the output may allow for better decisions to be made and give a clear avenue to appeal a decision made by an algorithm.

## 1.5 Responsibility and Accountability

There are numerous ways in which humans interact with algorithms within systems. Humans can act as supervisors – the algorithm makes initial decisions, but it is ultimately the human responsible for the system. Alternatively, humans may be oracles that the algorithm can query if some threshold (often based upon the uncertainty of a prediction) is met before the algorithm makes a final decision. The reverse scenarios are also possible, humans may be workers for algorithmic supervisors, or there may be an algorithm that a human can use as an oracle.

Human supervisors may like to trust the algorithm because it gives them some complex mathematical black box witchcraft that can be used to justify decisions without being accountable. The injustices caused by the COMPAS and PredPol algorithms are exacerbated by the problem that humans may just be passing the buck onto the algorithm to cover up their idleness or incompetence. When discussing COMPAS, Richard Berk, Emeritus Professor of Criminology and Statistics at the University of Pennsylvania, is quoted in Fry [19, p. 64] as saying “The courts are concerned about not making mistakes – especially the judges who are appointed by the public. The algorithm provides them with a way to do less work while not being accountable.” A review of PredPol by Lum and Isaac [56] noted that “Whereas before, a police chief could reasonably be expected to justify policing decisions, using a computer to allocate police attention shifts accountability from departmental decision-makers to black-box machinery that purports to be scientific, evidence-based and race-neutral.” As we have seen, the fact that the algorithms that are used are ‘scientific’ and ‘evidence-based’ by no means ensures that it will lead to accurate, reliable, or fair decisions [49].

This relationship needs to be adequately defined so that algorithms (and humans) can be held responsible when there are problems. There is also the question of who should have ultimate responsibility when an algorithm supervises human actions. This question is critical in high-risk situations, especially since supervising algorithms are not (yet) considered omniscient.

The Smiler is a rollercoaster at Alton Towers theme park in Staffordshire, UK. On 2nd June 2015, the ride malfunctioned, and engineers were called to come to fix the ride. After the fault was fixed, the engineers sent a test car onto the track. Meanwhile, the operators added an additional car to the track. Miscommunication between the engineers and the track operatives meant that the number of cars was miscounted, and the fact that the test car had stalled on track was missed. The safety algorithm present within the rides control systems did detect that there was a stalled car and triggered a block fault alarm. The engineers then overrode the algorithm and sent a car carrying 16 passengers around the track. The resulting crash with the stalled car caused five people serious injuries, including two people who required leg amputations [88].

The safety algorithm was supposed to supervise the rides running so that there were no incidents. The algorithm detects the movement of the cars by splitting the track into various blocks and having proximity switches that are activated as the car moves around the track, if the algorithm detects that a car has entered one zone but has not left, then it raises a block fault alarm. Aware of the potential fallibility of the system, the designers of the algorithm designed it to “err of the side of caution” “because [it] does not, and cannot, have eyes, the signals from the proximity switches are a proxy for reality, and may not always accurately reflect reality” [89, p. 17]. This erring of the side of caution caused numerous false alarms resulting in the operators believing that the algorithm was “crying wolf”, meaning that the operators ignored the error without justification. Preventing humane algorithms from crying wolf will be an important task.

The 737 MAX is the latest iteration of Boeing’s long-running 737 series of passenger aircraft. To fit the larger engines required by the 737 MAX, the engine position was moved forward slightly, which caused the aircraft’s flight characteristics to change, making it susceptible to stalling. So that pilots with licenses on other 737 aircraft were able to fly the new model with only limited training, Boeing added the Manoeuvring Characteristic Augmentation System (MCAS) to the aircraft so that the flight control system mimics the behaviour of the older generation. This algorithm adjusts the horizontal stabiliser trim of the aircraft to bring the aircraft nose down when it detects, using angle of attack (AoA) sensors, the aircraft may be approaching stall [90].

Two crashes have involved the 737 MAX, and in both cases, the MCAS system has been considered a major factor. On 29th October 2018, Lion Air Flight 610 crashed into the Java Sea 13 minutes after takeoff from Jakarta, Indonesia killing all 189 people on board [91]. Four months later, on 10th March 2019, Ethiopian Airlines Flight 302 crashed 6 minutes after takeoff from Addis Ababa, Ethiopia killing all 159 people on board [92]. In particular, the fact that the pilots flying the aircraft were unable to shut the system off completely meant that once the sensors started getting incorrect readings, the algorithm never ceded control [93]. Boeing implicitly decided that the algorithm would never make an error, despite relying on a single AoA sensor, whereas the humans might fail to fly the aircraft safely. This is clearly an inhumane policy.

The Smiler crash was, in part, caused by the algorithm ceding control to the human operators and human error causing disaster. Similarly, the crash of Air France 447 (discussed in Section 1.2) was caused by the algorithm giving up control and human error causing disaster. In both situations, the designers decided that the ultimate responsibility should lie with the humans. Whereas the 737 MAX accidents were caused by the algorithm not giving up control even when making a mistake.

The AF447 incident highlights another problem that can occur when algorithms become commonplace. The pilots were not used to flying by hand at night at altitude because they never needed to – the autopilot just did it – and they became reliant on it. Expecting the overseeing humans to step in and 'save the day' can be dangerous, not least as there is the possibility that they have forgotten how to do the task in the first place, a problem that Bainbridge [94] called *the irony of automation*.

The transition to self-driving cars will likely be littered with similar incidents. SAE International has defined six levels, from 0 as full human control to 5 being full automation in all possible conditions [95]. Levels 0-2 have the human driving with increasing help from the algorithm, and levels 4-5 have the algorithm drive the car with little to no assistance from the humans. At level 3, the algorithm controls the vehicle with the expectation that the human driver will respond appropriately to a request to intervene. This comes with the risks of similar incidents to The Smiler and AF447. There is also the problem that humans may not be paying enough attention. For instance, in 2018, Uber was testing a self-driving car when it crashed and killed a pedestrian, the human who was supposed to be supervising the system was looking at her phone [96].

An ethical question that is often proposed concerning self-driving cars is a restatement of the classical trolley problem [97]. If a self-driving car is travelling along a road and a tree falls over in front of it, a collision is inevitable, thus risking the lives of the passengers. The car could swerve to avoid the crash, but doing so would endanger the lives of pedestrians nearby. In essence, should your car kill a pedestrian or let you die in a crash? Philosophers have considered this problem extensively since it was first proposed [98–104].

The traditional trolley problem assumes a level of certainty that would not be possible in real life. There are numerous epistemic uncertainties involved in the situation that the car might have to consider. What if the car thinks it can slow down enough to lower the probability of a fatal accident for the pedestrian it will hit? The age of the passengers and pedestrian affects the risks of fatal injury in a collision; should the car try to estimate these whilst working out whether or not to swerve? Should the car know the number of passengers inside it, and if so, should that affect the decision-making process? It is possible that the passengers in the car, or pedestrians, might not die in the collision. All of these different uncertainties can lead to people making different decisions [103].

Some have argued that this uncertainty means that the trolley problem is irrelevant in the context of autonomous vehicles (see Himmelreich [104] as an example). They argue that there is so much epistemic uncertainty that it is impossible to discuss the problem reasonably. However, the nuances caused by the uncertainties involved mean that the problem becomes more interesting and relevant, not less. Just because trolley problems cannot "handle the problem of garden-variety risk" [102] does not mean that it is not a relevant thought experiment. Uncertainty matters when slight changes in initial conditions mean that one would make a different decision.

Such ethical discussions of the trolley problem assume that it will be a human that will programme what to do in such a situation. But, given that black-box machine learning algorithms will undoubtedly form the basis of many autonomous vehicles' self-driving elements, it is likely that no programmer will sit down and tell the algorithms to make such a decision. There is a question about who would be ultimately responsible for the decision: the human controller of the car? The company that made the car? The mathematics within the black box algorithm?

These questions will almost certainly be mediated, if not entirely decided by legislation [105]. There is also the problem that we may have entered an era where algorithms have gotten so complex that regulators and experts may not be able to certify them independently. This might lead to a situation where there may well be algorithms certifying algorithms. Therefore, it is vital to understand the risk associated with the decisions that the algorithms will make. Correctly understanding the uncertainty both into and out of the algorithm will be necessary to be able to deal with these risks [106].

## 1.6 Privacy

According to the Australian Bureau of Statistics [107] review of the 2011 Census, the average Australian is a married 37-year-old woman, born in Australia to parents also both born in Australia. She has English, Australian, Irish, or Scottish ancestry. She speaks only English at home and belongs to a Christian religion, most likely Catholic. She is married and lives with her husband and two children in a detached house with three bedrooms and two cars in a suburb of one of Australia's capital cities. Despite these broad sounding characteristics, none of the 22 million surveyed people met all of these characteristics. Such diversity among

Australians presents a problem when it comes to privacy.

Increasingly vast amounts of personal information are being collected about people every day. This data is used for both the greater good and less honourable reasons, and promises to revolutionise many industries such as medical science [108], advertising [109] and electioneering [110]. Data is collected, processed and sold with or without the consent of those included in the datasets. Many peer-reviewed journals require authors to share datasets used within their analysis in order to aid transparency and reproducibility [111].

Datasets that contain personal data are often anonymised by removing data thought to identify an individual (name, address, date of birth). There is, however, a re-identification tit for every de-identification tat [112]. This is often because defining a few characteristics can lead to very few matching individuals, like the average Australian. For instance, Rocher et al. [113] have shown that using only 15 demographic values, we can uniquely identify 99.98% of all Americans, and only gender, ZIP code and full date of birth can identify around 63% of the US population [114].

To illustrate, one could imagine a study that looks at how likely an individual has received a COVID vaccine based on socioeconomic status. It is not inconceivable that such a dataset may contain 100,000 individuals and include personal information such as their age, gender, race, education background and the district they live in. If Joe Bloggs agreed to take part in the study. He a 35-year-old white male lives who is university educated, lives in city-centre Liverpool and earns £33,000 per year. It may be possible to uniquely identify Joe even if the data has been de-identified – this is because there is unlikely to be that many other individuals that match Joe's characteristics.

Uncertainty can play a crucial part in helping protect privacy. If some of Joe's data was intervalised, it could help protect his privacy. For example, if his data was stored as being 35 to 40 white male who lives in a city in North West England earning £30-35,000 per year. Many more individuals could meet these criteria. Therefore, finding any particular individual within the data set is much more unlikely. Such internalisation can be done in such as way as not to lose the statistical information about the data set [115–117].

## 1.7 Discussion

Insofar as is possible, humane algorithms need to:

1. Accept control from humans, or relinquish it back to them, in ways that humans find workable,
2. Check human inputs for errors and misconceptions,

3. Self-diagnos problems or aberrant behaviour of the algorithms,
4. Protect privacy by securing or progressively anonymising personal information,
5. Create outcomes and situations that humans would judge as equitable or fair,
6. Recognise, accept and accommodate diversity among users,
7. Flexibly accept inputs from humans in disparate formats,
8. Does not require humans to respond to queries precisely, immediately, or at all,
9. Handle errors and unusual conditions in ways that do not result in catastrophic consequences,
10. Make conservative/fail-safe assumptions,
11. Do not unnecessarily burden humans,
12. Be transparent, or at least interrogatable, about its internal functioning,
13. Help humans to understand outputs and outcomes effected by the algorithms, and
14. Complement human skills.

Obviously, solving these problems and enabling algorithms to be humane is beyond the scope of a single thesis. There are, however, a number of the raised issues that this thesis will explore solutions to. All of these issues are associated with the theme of enabling algorithms to compute with imprecise information.

The first issue that is apparant from several of the examples is that uncertainties are often ignored or assumed away often as a result of algorithms inability to handle them directly. For instance, logistic regression is a popular supervised learning method used within a wide variety of studies, including the COMPAS algorithm above. Within traditional logistic regression analyses uncertainties are often reduced to a single middle-of-the-road model value. Chapter 5 considers how this uncertainty can be included within logistic regression models.

There is a difference between not knowing information and knowing that we do not know information. For instance, in the case of uncertainty bias knowing that we do not have a lot of information about a particular group of people may help alleviate against this bias. Similarly, if the PredPol algorithm could preceive that it did not know about crime rates where police officer were not on the beat and not implicity assume that there was no crime there it may have produced better results. Part of problem is that the mathematical methods used to express the epistemic uncertainty do not account for the difference between the epistemic and aleatory uncertainty. Chapter 2 will explore this difference, with Section 2.6.2 discussing it in detail.

As discussed in Section 1.1, inputs to algorithms need to be able to handle uncertain numbers. Often it is the case that computer models do not include these uncertainties intrusively and instead often rely on Monte Carlo methods to quantify the uncertainty. Chapter 3 will consider the creation of tools that can automatically include intrusive uncertainty analysis directly within their code.

Another issue that has been highlighted is the fact that many of the classification algorithms used rely on flawed information to assess their performance. Quantifying the uncertainty that is associated with this imprecision would be of great benefit for analysis using imperfect data. This problem is explored in Chapter 4. The problem if incorrect classifications within

# Chapter 2

# Probability Bounds Analysis

***Preface*** This section is an expanded version of Nicholas Gray, Scott Ferson, Marco De Angelis, Ander Gray, and Francis Baumont de Oliveira. Probability bounds analysis for Python. *Software Impacts*, 12:100246, May 2022. ISSN 26659638. doi: 10.1016/j.simpa.2022.100246.

This work lead to the creation of an open-source PBA library for Python, which makes use of SciPy [118], NumPy [119] and Matplotlib [120] in order to define, store, display and perform calculations with p-boxes and intervals within PBA [6].

Before the creation of this library, there was numerous PBA library for Python. Although versions did exist for Risk Calc [121], MATLAB [122], R [123] and Julia [124, 125], as Python is one of the most popular programming languages [126, 127], especially within the field of scientific computing, many scientists and engineers who prefer to programme using Python were unable to make use of the powerful methodology and many advantages of probability bounds analysis. The creation of PBA for Python expands the reach of this methodology so that it can be applied to other disciplines and sectors.

As discussed in Chapter 1, Section 1.1, two types of uncertainty, *aleatory* and *epistemic*, appear in the numerical calculations essential to science and engineering. Aleatory uncertainty arises from the natural variability in dynamical environments and material properties, errors in manufacturing processes or inconsistencies in the realisation of systems. Aleatory uncertainty cannot generally be reduced by empirical effort. Epistemic uncertainty is caused by measurement imperfections or a lack of understanding about the underlying physics or biology of a system. This could be due to not knowing the full specification of a system in the early phases of engineering design or simplifying the mathematics of a simulation to save computational resources.

Probability bounds analysis (PBA) is a tool that can be used to compute with both types of uncertainties without requiring often untenable assumptions to be made about the parameters involved in calculations and any subsequent dependencies between them. Probability bounds analysis has many applications across diverse disciplines ranging from aerospace engineering [128] to conservation biology [129]. The Wikipedia page lists many applications to various scientific problems*. It is particularly popular when undertaking risk or reliability analyses when data is not perfectly known [130–132]. PBA objects and methods can also be used within machine learning techniques [133, 134]. In this chapter, we discuss the fundamental components of PBA and how calculations are performed with them.

There are three main objects used for PBA, intervals, probability distributions and probability boxes (p-boxes). An interval is a value that is imprecisely known even though it may be fixed and unchanging, or perhaps an uncertain number representing values obeying an unknown distribution prescribed only by a specified range [135–139]. Intervals allow for epistemic uncertainty to be propagated through calculations.

A p-box is a generalisation of intervals and probability distributions in a single structure that allows the propagation of both epistemic and aleatory uncertainty through calculations in a rigorous way. A p-box can be considered as interval bounds on a probability distribution [140–142]. Within PBA it is convenient to think of a probability distribution as a special case of a p-box with precise parameterisation. Calculations performed with p-boxes yield results that are guaranteed to enclose all possible distributions of the output variable if the input p-boxes were also sure to enclose their respective distributions. The results may be best-possible if they cannot be made narrower without excluding valid distributions. Output p-box may also contain distributions that could not arise under any dependence between the two input distributions. This property allows them to be used for automatic verification of computer codes [142, 143].

As intervals embody epistemic uncertainty and probability distributions represent aleatory uncertainty, p-boxes can be used to propagate both type of uncertainty through calculations. Libraries that add some or all of these objects are available in C++ [144], MATLAB [122], R [123 or 145], Julia [125] and the Python library I created [6].

There are other objects that could be used within uncertainty analysis that all have nu-

*https://en.wikipedia.org/wiki/Applications_of_p-boxes_and_probability_bounds_analysis merous applications within engineering [132], including, but not limited to: second-order distributions or meta-distributions [146, 147], fuzzy numbers [148], possibility distributions [149], consonant structures [150] and info-gap models [151]. Only PBA will be considered within this thesis.

## 2.1 Intervals

An unknown real number $x$ can be represented by an interval $[\underline{x}, \overline{x}]$, where $\underline{x} \leq x \leq \overline{x}$. This implies that the true value of $x$, $x^{\dagger}$, can be any number within the range. Intervals do not make any futher assumptions about which values within the range are more or less likely than other values. In addition to expressing the intervals using the lower and upper bound of the range, they can also be expressed $[x_m \pm \Delta]$, where $x_m$ is the midpoint of the interval and $\Delta$ is the half-width of the interval.

Within the context of probability bounds analysis, it is useful to consider intervals as the set of all possible probability distributions that lie within the endpoints of the interval, again, with no individual distribution considered more or less accurate. This definition is discussed further in Section 2.4. A zero-variance interval is an interval that is known or assumed to have a single, fixed value in the range.

If $a = [\underline{a}, \overline{a}]$ and $b = \left[\underline{b}, \overline{b}\right]$ are intervals, then the following arithmetic operations can be performed in PBA:

- **Addition**

$$a + b = [\underline{a} + \underline{b}, \overline{a} + \overline{b}] \tag{2.1}$$

- **Subtraction**

$$a - b = [\underline{a} - \overline{b}, \overline{a} - \underline{b}] \tag{2.2}$$

- **Multiplication**

$$a * b = \left[\min\left(\underline{a} * \underline{b}, \underline{a} * \overline{b}, \overline{a} * \underline{b}, \overline{a} * \overline{b}\right), \max\left(\underline{a} * \underline{b}, \underline{a} * \overline{b}, \overline{a} * \underline{b}, \overline{a} * \overline{b}\right)\right] \tag{2.3}$$

- **Division**

$$\frac{a}{b} = a * \left[\frac{1}{\overline{b}}, \frac{1}{\underline{b}}\right] \tag{2.4}$$

If $0 \in b$ then $^{a}/_{b}$ returns a division-by-zero error. If there is dependence between two intervals then PBA allows for this dependence to be included within the calculation. For intervals, perfect and opposite dependence calculations are possible. Perfect dependence between $a$

and $b$ implies that larger values of $a$ correspond to larger values of $b$. In this scenario the arithmetic operations become:

$$a \circ b = \left[\underline{a} \circ \underline{b}, \overline{a} \circ \overline{b}\right] \tag{2.5}$$

where $\circ \in (+, -, *, /)$, whereas, under opposite dependence smaller values of $a$ imply larger values of $b$, meaning that the arithmetic operations are:

$$a \circ b = \left[\underline{a} \circ \overline{b}, \overline{a} \circ \underline{b}\right]. \tag{2.6}$$

An interval can be propagated through a function producing an interval output, $f\left([\underline{x}, \overline{x}]\right) = \left[\underline{y}, \overline{y}\right]$ where $\underline{y}$ is the minimum possible value of $f(x)$ for all $x \in [\underline{x}, \overline{x}]$ and $\overline{y}$ is the maximum possible value. This calculation is simple for monotonic functions. For instance, increasing monotonicity implies that the end points of the input interval correspond to the end points of the output interval, i.e.

$$f([\underline{a}, \overline{a}]) = [f(\underline{a}), f(\overline{a})]. \tag{2.7}$$

For more general functions alternative strategies are needed to ensure correct calculations [152].

Comparison operations can be performed on intervals, however, the uncertainty associated with the interval leads to uncertainty in the Boolean operations. For example, if a decision relies on some value $x$ being less than a fixed value $a$, when we know the value of $x$ accurately then it is easy to make such a comparison. However, if there is some uncertainty about the value of $x$ then this comparison may not be so easy. The comparison becomes

$$x < a = \begin{cases} 1 & \text{if } \overline{x} < a \\ 0 & \text{if } \underline{x} \geq a \\ [0, 1] & \text{otherwise} \end{cases} \tag{2.8}$$

with 0 and 1 denoting false and true respectively, and [0,1] being the Boolean equivalent of “I don’t know”. We can call [0,1] the *dunno* interval. Similarly,

$$x > a = \begin{cases} 1 & \text{if } \underline{x} > a \\ 0 & \text{if } \overline{x} \leq a \\ [0, 1] & \text{otherwise.} \end{cases} \tag{2.9}$$

For intervals it is impossible to say whether an interval is equal to a precise value,

$$x == a = \begin{cases} [0,1] & \text{if } a \in x \\ 0 & \text{otherwise.} \end{cases} \tag{2.10}$$

Two intervals can also be compared to each other. For intervals $x = [\underline{x}, \overline{x}]$ and $y = [\underline{y}, \overline{y}]$, then

$$x < y = \begin{cases} 1 & \text{if } \overline{x} < \underline{y} \\ 0 & \text{if } \underline{x} \geq \overline{y} \\ [0,1] & \text{otherwise} \end{cases} \tag{2.11}$$

and

$$x > y = \begin{cases} 0 & \text{if } \overline{x} \leq \underline{y} \\ 1 & \text{if } \underline{x} > \overline{y} \\ [0,1] & \text{otherwise.} \end{cases} \tag{2.12}$$

This implies that we cannot say whether an uncertain value characterised by an interval is larger or smaller than another unless the interval is entirely greater or less than the other interval. For the equality comparison,

$$x == y = \begin{cases} [0,1] & \text{if } x \cup y \neq \varnothing \\ 0 & \text{otherwise,} \end{cases} \tag{2.13}$$

it is never possible to say that one value is equal to another. We can introduce a new Boolean operator (===) to test whether two uncertain numbers are equivalent in form,

$$x === y = \begin{cases} 1 & \text{if } \underline{x} = \underline{y} \text{ and } \overline{x} = \overline{y} \\ 0 & \text{otherwise.} \end{cases} \tag{2.14}$$

The dunno interval, $[0,1]$, can be converted into a Boolean value using adverb operators such as ALWAYS or SOMETIMES,

$$\text{ALWAYS}\,([0,1]) = 0 \tag{2.15a}$$

$$\text{SOMETIMES}\,([0,1]) = 1 \tag{2.15b}$$

so that we can get

$$\text{ALWAYS}\,(x < y) = \begin{cases} 1 & \overline{x} < \underline{y} \\ 0 & \text{otherwise} \end{cases} \tag{2.16}$$

$$\text{SOMETIMES}\,(x < y) = \begin{cases} 1 & \underline{x} < \overline{y} \\ 0 & \text{otherwise.} \end{cases} \tag{2.17}$$

## 2.2 Probability Distributions and Probability Boxes

A probability distribution is a mathematical function that gives the probabilities of occurrence for different possible values of a variable. Probability boxes (p-boxes) represent interval bounds on probability distributions. The simplest kind of p-box can be expressed mathematically as

$$\mathcal{F}(x) = [\underline{F}(x), \overline{F}(x)],\ \underline{F}(x) \geq \overline{F}(x)\ \forall x \in \mathbb{R} \tag{2.18}$$

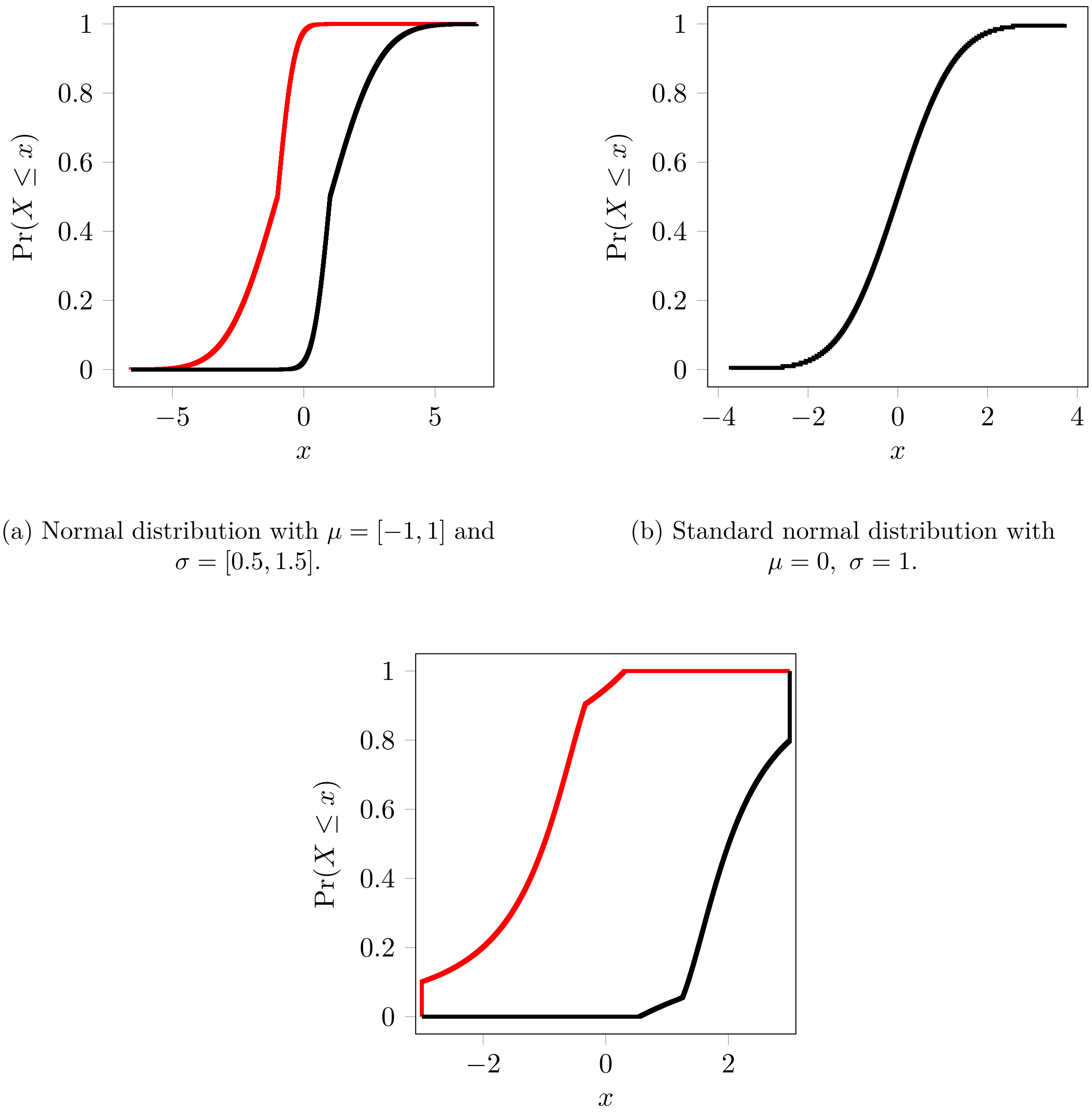
where $\underline{F}(x)$ is the function that defines the left bound of the p-box and $\overline{F}(x)$ defines the right bound of the p-box. Figure 2.1a shows a p-box that is defined by a normal distribution with $\mu = [-1, 1]$ and $\sigma = [0.5, 1.5]$. In this plot $\underline{F}(x)$ is shown in red and $\underline{F}(x)$ is shown in black.h

Naturally, precise probability distributions can be specified in PBA by defining a p-box with precise parameters. This means that within PBA probability distributions are considered a special case of a p-box with zero width. Consequently, all methodology that applies to p-boxes can also be applied to probability distributions. Figure 2.1b shows a standard normal distribution ($\mu = 0$, $\sigma = 1$).

Distribution-free p-boxes can also be generated when the underlying distribution is unknown but parameters such as the mean, variance or minimum/maximum bounds are known. Such p-boxes make no assumption about the shape of the distribution and instead return bounds enclosing all possible distributions that are permissible given the known information. Such p-boxes can be constructed making use of Chebyshev, Markov and Cantelli inequalities from probability theory. A p-box defined with min $= -3$ ,max $= 3$, $\mu = [0, 1]$ and $\sigma = 1$ is shown in Figure 2.1c.

As with intervals, standard arithmetic operations can be performed on p-boxes (and therefore probability distributions which are special cases of p-boxes). For two p-boxes $\mathcal{A}(x) = [\underline{A}(x), \overline{A}(x)]$ and $\mathcal{B}(x) = [\underline{B}(x), \overline{B}(x)]$,

$$\mathcal{C}(x) = \mathcal{A}(x) \circ \mathcal{B}(x) = [\underline{C}(x), \overline{C}(x)] \tag{2.19}$$



(a) Normal distribution with $\mu = [-1, 1]$ and $\sigma = [0.5, 1.5]$.

(b) Standard normal distribution with $\mu = 0,\ \sigma = 1$.

(c) Distribution-free p-box with min $= -3$ ,max $= 3$,$\mu = [0, 1]$ and $\sigma = 1$

Figure 2.1: Probability distributions and probability boxes.

where

$$\underline{C}(z) = \inf_{z=x\circ y} \left[\min\left(\underline{A}(x) \circ \underline{B}(y), 1\right)\right] \tag{2.20a}$$

$$\overline{C}(z) = \sup_{z=x\circ y} \left[\max\left(\overline{A}(x) \circ \overline{B}(y) - 1, 0\right)\right] \tag{2.20b}$$

if $\circ \in [+, \times]$, or

$$\underline{C}(z) = 1 + \inf_{z=x\circ y} \left[\min\left(\underline{A}(x) \circ \overline{B}(y), 0\right)\right] \tag{2.21a}$$

$$\overline{C}(z) = \sup_{z=x\circ y} \left[\max\left(\overline{A}(x) \circ \underline{B}(y), 0\right)\right] \tag{2.21b}$$

if $\circ \in [-, \div]$ [141, p. 89]. If $0 \in \mathcal{B}$ then the division returns a error.

## 2.3 Dependencies and the Repeated Variable Problem

When performing uncertainty analysis it would be ideal to always obtain best possible results that are guaranteed to bound the true value without overestimating the uncertainty. The uncertainty can be inflated or artifactually high if careful consideration of the dependence between, and repetition of, uncertain numbers is not undertaken.[†]

For example, if $a = [1, 2]$, $b = [-1, 1]$ and $c = [3, 4]$, then

$$\begin{aligned} ab + ac &= [1,2] \times [-1,1] + [1,2] \times [3,4] \\ &= [-2,2] + [3,8] \\ &= [1,10] \end{aligned} \tag{2.22}$$

but

$$\begin{aligned} a(b + c) &= [1,2]\left([-1,1] + [3,4]\right) \\ &= [1,2] \times [2,5] \\ &= [2,10]. \end{aligned} \tag{2.23}$$

Although algebraically these two expressions are equal, the uncertainty associated with $ab + ac$ is greater than the uncertainty about $a(b + c)$. This is because the uncertain variable $a$ is repeated within the former but appears only once in the latter. In essence, the uncertainty about $a$ has been considered twice when performing the first calculation. The amount of this artifactual uncertainty can be reduced by transforming the original

[†] *Artifactual uncertainty* is when the size of the outputted uncertainty is overly wide as an artifact of the calculationand not as a result of the true uncertainty present.

equation into a single-use expression where uncertain variables are only used once. If this is not possible, other techniques can be used to reduce this artifactual uncertainty (e.g. [153–156]). The repeated variable problem appears to be ubiquitous to many, if not all, uncertainty calculi [132].

Knowledge of what the dependence is between the two p-boxes can reduce the amount of uncertainty present within the output p-box. Figure 2.2 shows the result of adding a normal p-box $\mathcal{A} = \mathrm{N}([-1,1],1)$ to a uniform p-box $\mathcal{B} = \mathrm{U}([0,1],[2,3])$, with different dependencies between $\mathcal{A}$ and $\mathcal{B}$. When the dependence between $\mathcal{A}$ and $\mathcal{B}$ is unknown, the operation defined in Equations 2.19,2.20 and 2.21 yields the most general bounds guaranteed to enclose the true distribution of $\mathcal{C} = \mathcal{A} + \mathcal{B}$ which are called the Fréchet bounds. As depicted in Figure 2.2, the Fréchet bounds enclose all the other dependencies. Perfect (or comonotonic) dependence is where there is a perfect positive relationship between the two variables, with the highest possible correlation coefficient. Opposite (or countermonotonic) dependence creates a perfect negative relationship between the two variables with the lowest possible correlation coefficient. Independence is where there is no dependence between the two variables. It should not be assumed that variables are independent unless this is known because wrongly assuming independence can lead to incorrectly reducing the amount of uncertainty and understating tail risks.

## 2.4 Comparison Between Objects

As mentioned within Section 2.1, intervals can be considered as the set of all possible distributions that lie between the endpoints of the interval. This implies that interval objects can be converted into p-boxes by transforming the interval into a box-shaped p-box, such an object is shown with the black lines in Figure 2.3. This property means that arithmetic can be performed between p-boxes and intervals by casting the interval as a p-box when performing the calculation. Conversely, many unary operations that can be performed on intervals can be performed on p-boxes. This can be done by slicing the p-box into intervals, performing the operation before sorting and recombining the intervals back into a p-box.

Within PBA an interval can be considered as the most basic object, for example, if all we know about variable $x$ is that its value lies between -2 and 2 then all we can say is that $x = [-2, 2]$, this is shown as the black lines in Figure 2.3. If more information about the variable is known then the uncertainty can be reduced, for instance, if we know that $x$ has mean 0.5, then we can instead use a distribution-free p-box to model the uncertainty, this is shown with the blue lines. If we also know that the standard deviation is 1 we can

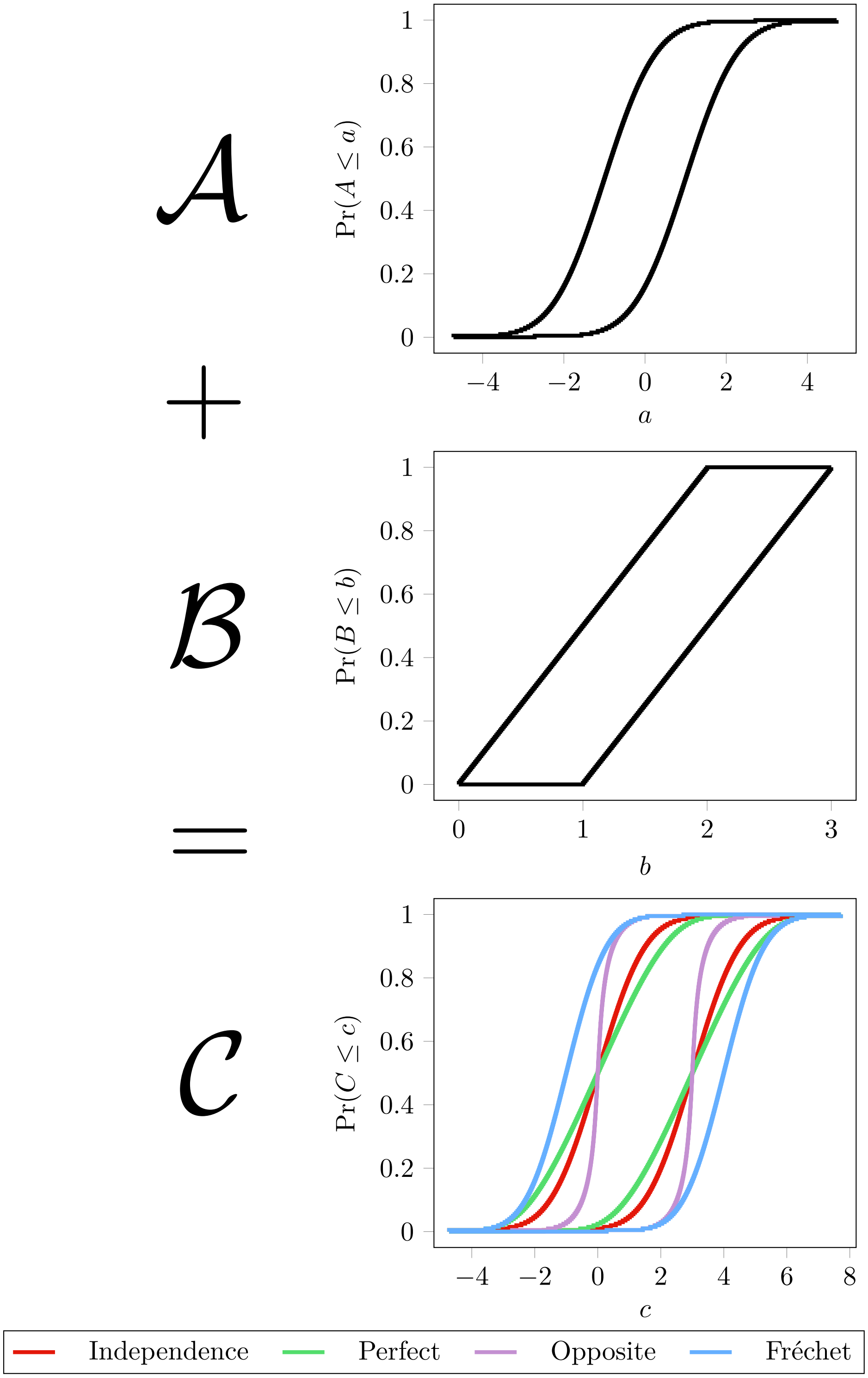


Figure 2.2: Adding together two p-boxes with different dependencies.

reduce the area of the p-box as shown with the red lines. Finally, if we know that $x$ follows a truncated normal distribution then we can model $x$ as shown with the green line. As calculations with all of these objects can be performed using PBA, analysts can compute with what they know rather than making assumptions that may be unjustified.

## 2.5 Confidence Boxes

Confidence boxes (c-boxes) are imprecise generalisations of traditional confidence distributions, which, like Student's t–distribution, encode frequentist confidence intervals for parameters of interest at every confidence level [157–159]. They are analogous to Bayesian posterior distributions in that they characterise the inferential uncertainty about distribution parameters estimated from sparse or imprecise sample data, but they have a purely frequentist interpretation that makes them useful in engineering because they offer a guarantee of statistical performance across repeated use. Unlike confidence intervals which traditionally cannot be used in mathematical calculations, c-boxes can be propagated through mathematical expressions using the ordinary machinery of probability bounds analysis, and this allows analysts to compute with confidence, both figuratively and literally, because the results also have the same confidence interpretation. For instance, they can be used to compute probability boxes for both prediction and tolerance distributions.

If $C(\theta, \mathbf{x}) = [\underline{C}(\theta, \mathbf{x}), \overline{C}(\theta, \mathbf{x})]$ is a c-box generated for an unknown parameter $\theta \in \Theta$ from a given dataset $\mathbf{x} \sim F(\theta_0)$. For a desired confidence level $\alpha$ a confidence interval $\boldsymbol{\alpha} = [\underline{\alpha}, \overline{\alpha}] \subseteq [0, 1]$ is created such that $\overline{\alpha} - \underline{\alpha} = \alpha$. An $\alpha$ confidence interval for $\theta_0$ can be found using

$$\boldsymbol{\theta}_\alpha \sim \left[\underline{\theta}, \overline{\theta}\right] = [\underline{C^{-1}}(\alpha, \mathbf{x}), \overline{C^{-1}}(\alpha, \mathbf{x})]. \tag{2.24}$$

Confidence intervals are usually either centered ($\boldsymbol{\alpha} = [\frac{1}{2} \pm \frac{\alpha}{2}]$) or one sided ($\boldsymbol{\alpha} = [0, \alpha]$ or $\boldsymbol{\alpha} = [1 - \alpha, 1]$).

For a c-box to be valid, it must have the confidence interpretation. The basic definition is that for $C(\cdot)$ to be a valid, all $100\alpha\%$ confidence intervals that are drawn from $C(\cdot)$ must contain the true parameter value at least $100\alpha\%$ of the time. A modern definition was composed by Schweder and Hjort [160] and Singh et al. [161] who define two requirements. For a function $C(\cdot) = C(, \cdot, \mathbf{X})$ to be a valid confidence distribution for $\theta$, then:

(i) for each $x_i \subseteq \mathbf{x}, C(\cdot) \equiv C(\cdot, \mathbf{x})$ is a continuous cumulative distribution on $\Theta$, and

(ii) at the true parameter value $\theta = \theta_0,\ C(\theta_0) \equiv C(\theta_0, \mathbf{x}_n)$ as a function of the sample $\mathbf{x}_n$ follows the uniform distribution $\mathrm{U}(0, 1)$.

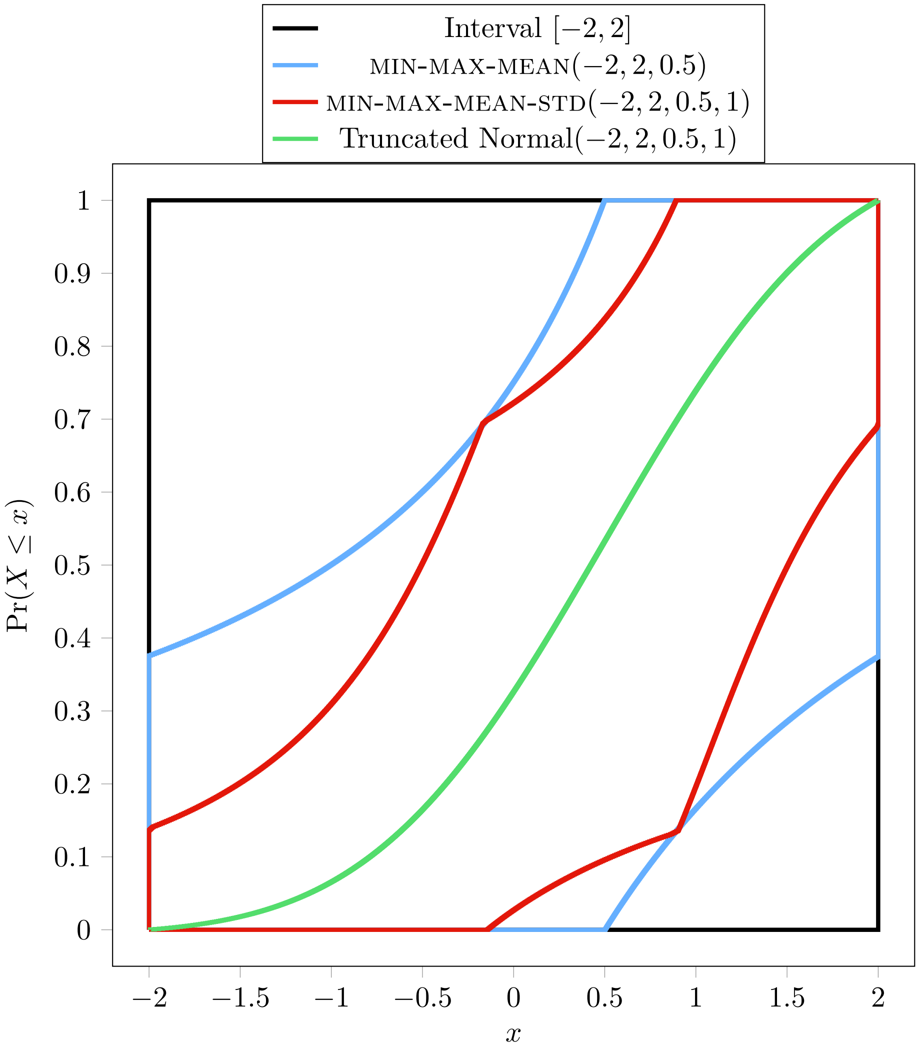


Figure 2.3: Comparing different PBA objects as the amount of information about $x$ increases. The interval contains less information about the value than the min-max-mean p-box which in turn has less information than the min-max-mean-standard-deviation p-box. The probability distribution with known shape contains the most information about $X$.

### 2.5.1 Singh Plots

This definition can be tested by making use of Singh plots which visualise the performance of a proposed structure in terms of both coverage and aspects of conservatism [7]. A Singh plot is generated by estimating the minimum coverage probability for a series of $\alpha$ levels, using the procedure shown in Algorithm 2.1.

**Algorithm 2.1:** Generation of a Singh plot. Taken from Wimbush et al. [7]

```
Input: $C^*$ ←Proposed confidence procedure
$F(\boldsymbol{\theta})$ ←Target distribution taking parameters $\boldsymbol{\theta}$
$\theta_0$ ←True value of parameter of interest
$m$ number of iterations
$n$ number of samples
for all $i \in 1, \ldots, m$ do
    Generate sample: $\boldsymbol{x} = \{x_1, \ldots, x_n\} \sim F(\boldsymbol{\theta})$;
    Calculate minimum required confidence for coverage: $\alpha_i = C^*(\theta_0, \boldsymbol{x})$
end
$S(\alpha) \leftarrow$ ECDF of $\alpha_1, \ldots, \alpha_m$, $S$
Plot $S(\alpha)$ and CDF of $U(0,1)$ for comparison
Output: Singh plot for visual assessment of confidence procedure properties
```

As imprecise confidence structures, a c-box will naturally produce upper and lower bounds rather than a single plot:

$$S(\alpha) = \left[\underline{S(\alpha)}, \overline{S(\alpha)}\right] \tag{2.25}$$

where

$$\underline{S}(\alpha) = \Pr(\underline{C}\,(\theta, \mathbf{x}) \geq \alpha) \tag{2.26}$$

and

$$\overline{S}(\alpha) = 1 - \Pr(\overline{C}\,(\theta, \mathbf{x}) \geq \alpha). \tag{2.27}$$

For visual clarity, the following can be used as an alternative expression for $\overline{S}(\alpha)$:

$$\overline{S}(\alpha) = \Pr(\overline{C}\,(\theta, \mathbf{x}) < \alpha) \tag{2.28}$$

Technically this does not portray the coverage property of the upper bound, as it effectively treats the upper bound as if it were a lower bound for drawing confidence intervals. However, it is generated from the same information and produces a more intuitive plot. As the intent of a Singh plot is to rapidly convey the coverage of a confidence structure, this is deemed to be a valuable property over technical accuracy. Here, deviations from the $U(0, 1)$ diagonal represent a number of limitations in the confidence structure used, including the finite samle of data on which that it is based, but so long as neither bound crosses this line then the coverage requirement is satisfied. This property allows Singh plots to rapidly convey the coverage characteristics of c-boxes.

### 2.5.2 Example

C-boxes can be computed in a variety of ways directly from random sample data. There are c-boxes both for parametric problems (where the family of the underlying distribution from which the data were randomly generated is known to be normal, lognormal, exponential, Bernoulli, Poisson, etc.), and for nonparametric problems in which the shape of the underlying distribution is unknown. C-boxes account for the uncertainty about a parameter that comes from the inference from observations, including the effect of small sample size, but also the effects of imprecision in the data and demographic uncertainty which arises from trying to characterise a continuous parameter from discrete data observations. Within this thesis we shall consider $k$-out-of-$n$ c-boxes, mathematically expressed as [158]:

$$\mathrm{KN}(x|k,n) = [\mathrm{beta}(k, n-k+1), \mathrm{beta}(k+1, n-k)] \tag{2.29}$$

These c-boxes express the uncertainty associated with the number of observations, $k$, of an event given that there has been $n$ trials. For example, a KN c-box with $k = 3$ and $n = 10$ is shown in Figure 2.4.

We can test whether Equation 2.29, does indeed return structures that have the confidence by testing using Singh plots. We can do this by selecting various different $n$ and $p$ values with $k \sim \mathrm{binomial}(p, n)$, then using Algorithm 2.1 to produce the Singh plots shown in Figure 2.5. These plots show that the method *singhs* (has the confidence interpretation) for all the shown $p$ and $n$. The width of the c-boxes around the diagonal represent the conservatism of the c-box.

An alternative approach is to vary the input values to produce a *global average* Singh plot (GA-Singh plot). For the KN c-box, this can be achieved by randomly sampling $p \in [0, 1]$ and $n \in [0, 10^5]$, then $k \in \mathrm{binomial}(p, n)$ for $10^4$ iterations and then using Algorithm 2.1 to produce the Singh plot shown in Figure 2.6. This shows that on average the method produces confidence structures that Singh on average. It is important to note that the GA-Singh plot does not prove that the method *always* produces admissable confidence structures. As with all averages it may hide some nefarious results, especially as it may be difficult to get to extreme values that are not likely to Singh using such an approach. However, a GA-Singh plot that does not Singh implies that the method does not have the confidence interpretation for the sampled values and therefore likely not be fit for purpose.

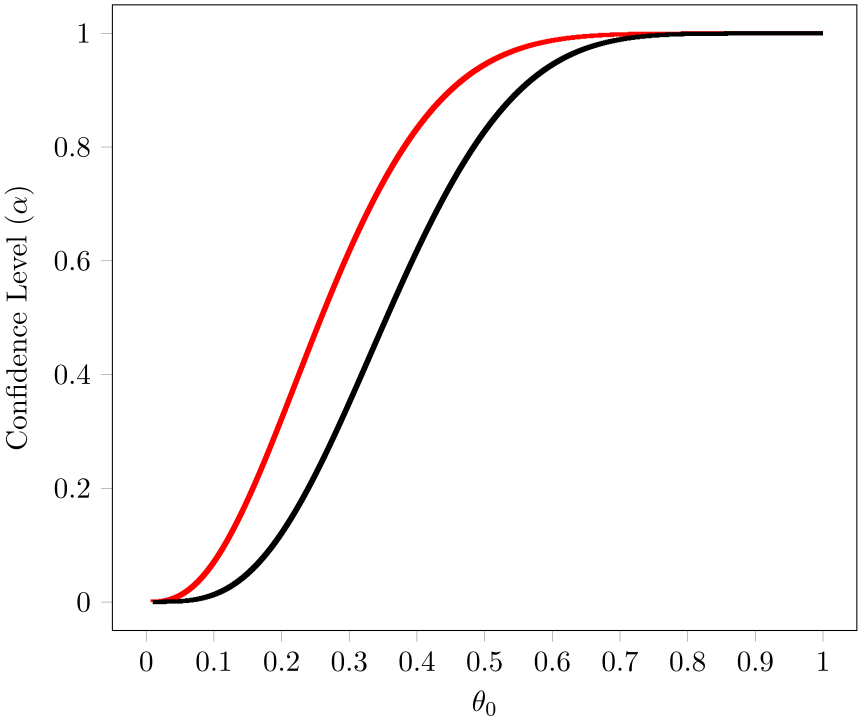


Figure 2.4: 3-out-of-10 confidence box for a binomial rate using Equation 2.29.

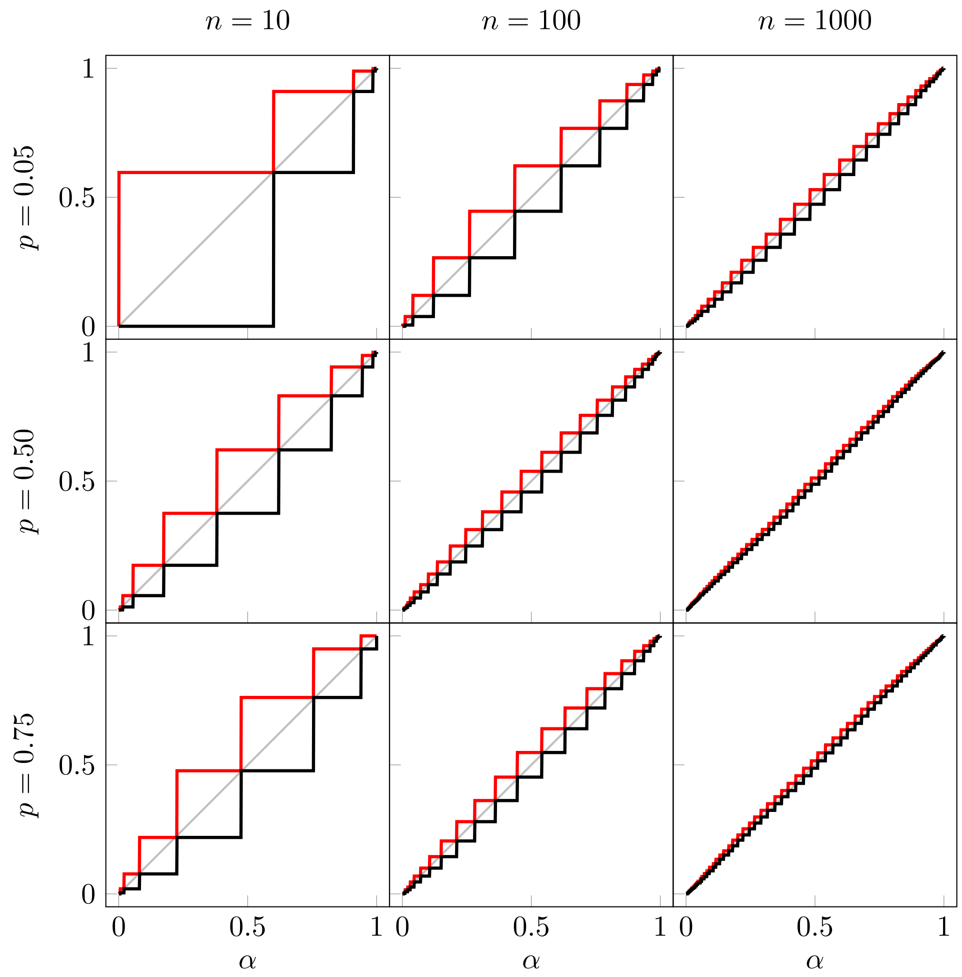


Figure 2.5: Singh plot for the $k$-out-of-$n$ c-box for various values of $n$ and $p$ with $k \sim \text{binomial}(p, n)$ for $10^4$ iterations.

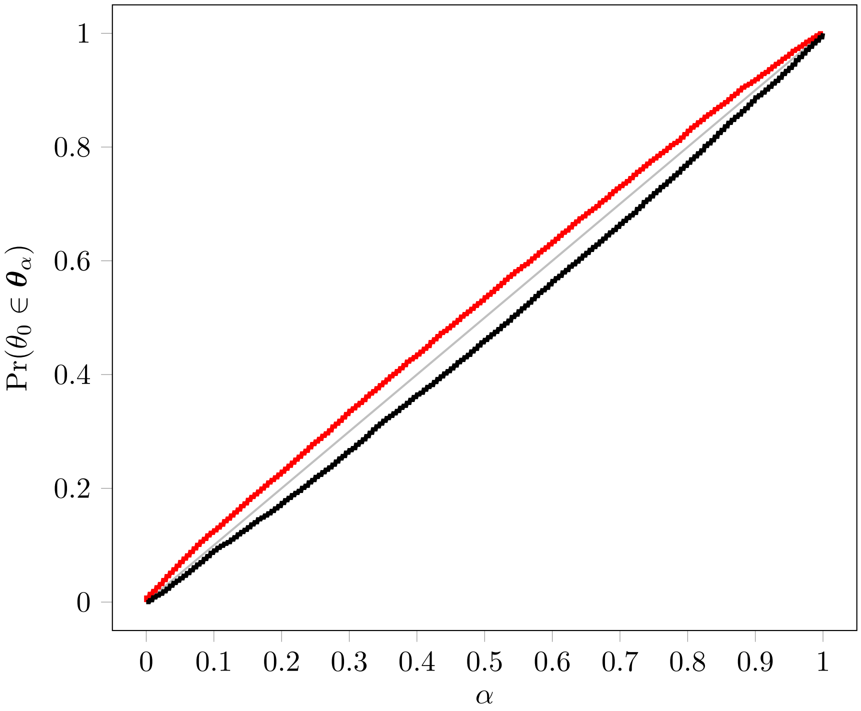


Figure 2.6: Global average Singh plot for the $k$-out-of-$n$ c-box with $10^4$ iterations.

## 2.6 Example Calculations

In order to demonstrate the usefulness of probability bounds analysis when performing caluclations we can consider two problems in aerospace engineering.

### 2.6.1 Attitude of a Satellite

The attitude of a spacecraft is the direction in which it points. It is often important to control the attitude of a satellite as its solar panels need to be pointed at the sun, communication antennas need to be pointed at the earth or scientific instruments need to point at the correct target. Attitude can be controlled through reaction wheels which can provide angular momentum to the satellite to point it in the desired direction.

The choice of how powerful a reaction wheel needs to be depends on the torque needed to change the attitude of the satellite. The torque required depends on the moment of inertia of the satellite. The moment of inertia depends on the size of the satellite's solar panels which impacts the power available to the reaction wheels, which impacts the torque available, and so on. Therefore, whilst there is uncertainty about the design of the satellite it is useful to make calculations using imprecise numbers. There are also additional uncertainties to consider such as the fact that solar radiation is not constant.

The equations of motion that determine the required angular momentum from the reaction wheel to change the attitude of a satellite within 1 dimension are as follows:

$$\tau_{slew} = \frac{4\theta_{\text{slew}}}{\Delta t^2_{\text{slew}}} I \tag{2.30}$$

$$\tau_{dist} = \tau_g + \tau_{\text{sp}} + \tau_m + \tau_a \tag{2.31}$$

$$\tau_g = \frac{3\mu}{2(R_E + H)^3} \left| I_{\max} + I_{\min} \right| \sin(2\theta) \tag{2.32}$$

$$\tau_{sp} = L_{sp} \frac{F_S}{c} A_s (1 + q) \cos(i) \tag{2.33}$$

$$\tau_m = \frac{2MD\mu_0}{(R_E + H)^3} \tag{2.34}$$

$$\tau_a = \frac{1}{2} L_a \rho C_d A V^2 \tag{2.35}$$

$$V = \sqrt{\frac{m}{R_E + H}} \tag{2.36}$$

$$h = \Delta t_{\text{orbit}} (\tau_{\text{slew}} + \tau_{\text{dist}}) \tag{2.37}$$

Table 2.1 gives definitions and values for all variables within these equations.

PBA for Python can be used to perform the calculation[‡] using the uncertainty expressed about the variables in Table 2.1. Figure 2.7 shows the final step in the calculation (Equation 2.37). The resultant p-box can be used to make decisions about the requirements of the reaction wheels.

### 2.6.2 Satellites Crashing

Balch et al. [58] consider two satellites are on orbital trajectories that will pass close to each other. How confident can we be that they will not collide?They consider this question within the complicated dynamics of two satellites in orbit. However, we can simplify the mathematics of the scenario and still demonstrate the dynamics presented in the original paper. To simplify the dynamics, we can treat one satellite as a stationary target with a particle (the other satellite) moving towards it.

Let there be a target of width $d$ and the centre of the target at the origin in line with $y = 0$ and the particle starts at $\mathbf{x}(0) = (x_0, y_0)$ moving with velocity $\mathbf{x}'(\mathbf{t}) = (x'(t), y'(t))$, as shown in Figure 2.8. If at time $T$ the particle is at $(r, 0)$, then the particle will hit the target if

$$|r| \leq \frac{d}{2}. \tag{2.38}$$

Deterministically we can find $r$ by solving the equations of motion, finding $T$ and subsequently $r$. For example, if the particle starts at $\mathbf{x}(0) = (3, 4)$ with velocity $\mathbf{x}'(t) = (-t/2, -1)$, then we can solve the equations of motion to get the co-ordinates at time $t$,

$$\mathbf{x}(t) = \left( -\frac{-t^2}{4} + 3, -t + 4 \right). \tag{2.39}$$

[‡] The code to perform this calculation can be found at https://codeocean.com/capsule/8485409/tree/v2/code/attitude.ipynb.

| **Symbol** | **Variable** | **Type** | **Value** | **Units** |
|---|---|---|---|---|
| $h$ | Required angular momentum | | Calculated | N m s |
| $\tau_{\text{tot}}$ | Total required torque | | Calculated | N m |
| $\tau_{\text{slew}}$ | Slewing torque | | Calculated | N m |
| $\tau_{\text{a}}$ | Torque due to atmospheric resistance | | Calculated | N m |
| $\tau_{\text{sp}}$ | Torque due to solar radiation pressure | | Calculated | N m |
| $\tau_{\text{g}}$ | Torque due to gravitational gradient | | Calculated | N m |
| $V$ | Velocity of satellite | | Calculated | m s$^{-1}$ |
| $C_d$ | Drag coefficient | p-box | min=2, max = 4, mean=3.13 | unitless |
| $L_a$ | Aerodynamic drag torque moment | p-box | min=0, max=3.75, mean=0.25 | m |
| $L_{\text{sp}}$ | Solar radiation torque moment | p-box | min=0, max=3.75, mean=0.25 | m |
| $D$ | Residual dipole | interval | [0,1] | A m$^2$ |
| $i$ | Sun incidence angle | interval | [0,90] | degrees |
| $\rho$ | Atmospheric density | interval | $[3.96, 99] \times 10^{-12}$ | kg m$^3$ |
| $\theta$ | Major moment axis deviation from nadir | interval | [10,19] | degrees |
| $q$ | Surface reflectivity | interval | [0.1,0.99] | unitless |
| $I_{\text{min}}$ | Minimum moment of inertia | point | 4655 | kg m$^2$ |
| $I_{\text{max}}$ | Maximum moment of inertia | point | 7315 | kg m$^2$ |
| $m$ | Earth gravity constant | point | $3.98{\times}10^{14}$ | m$^3$s$^{-2}$ |
| $A$ | Area in the direction of flight | point | 3.752 | m$^2$ |
| $R_E$ | Earth radius | point | 6378.14 | km |
| $H$ | Orbit altitude | point | 340 | km |
| $F_S$ | Average solar flux | point | 1367 | W m$^{-2}$ |
| $q_{\text{slew}}$ | Maximum slewing angle | point | 38 | degrees |
| $c$ | Speed of light | point | $2.9979{\times}10^8$ | m s$^{-1}$ |
| $M$ | Earth magnetic moment | point | $7.96{\times}10^{22}$ | A m$^2$ |
| $\Delta t_{\text{slew}}$ | Minimum manoeuvre time | point | 760 | s |
| $A_s$ | Area reflecting solar radiation | point | $3.75^2$ | m$^2$ |
| $\Delta t_{\text{orbit}}$ | Quarter orbit period | point | 1370 | s |
| $\mu_0$ | Permiability of free space | point | $4\pi \times 10^{-7}$ | N A$^{-2}$ |

Table 2.1: Definitions and values for Equations 2.30–2.37

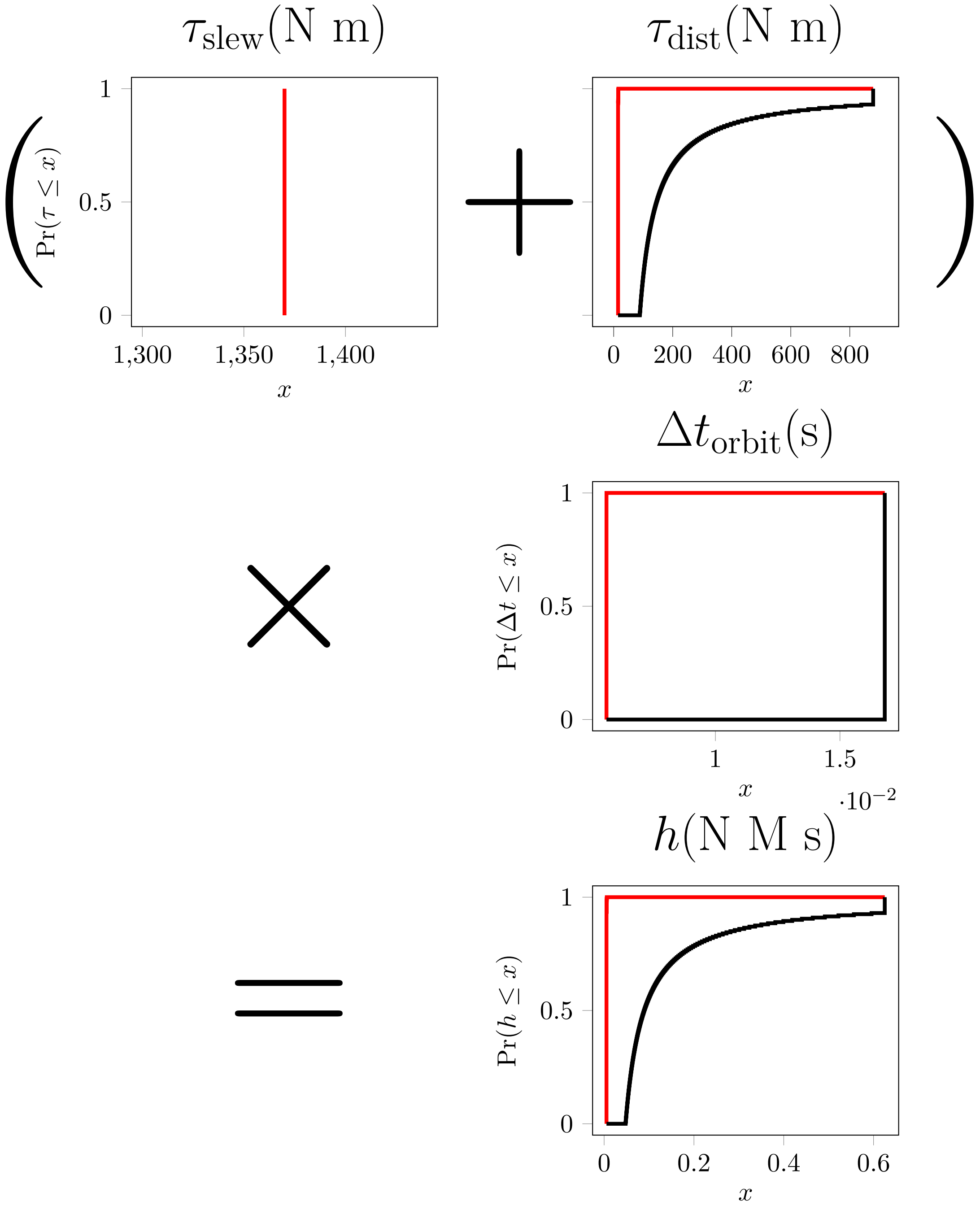


Figure 2.7: Calculation of Equation 2.37 using probability bounds analysis.

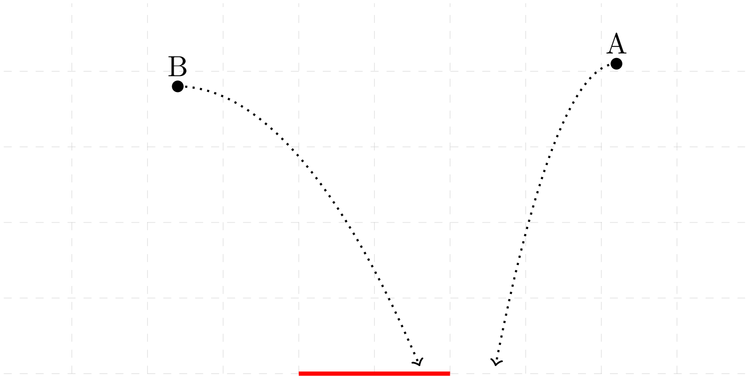


Figure 2.8: Will the particle hit the target?

This equation can then be used to calculate that $T = 4$ and, therefore, $r = -7$, meaning we can conclude the particle will miss the target.

However, if there is uncertainty about $\mathbf{x}$ or $\mathbf{x}'$, then it may only be possible to get an estimate, $\hat{r}$, for $r$. We would have to consider the probability, $p$, that the objects collide,

$$p = \Pr(|r| \leq {}^{d}\!/_{2}). \tag{2.40}$$

If we assume that $r$ is normally distributed with mean $\hat{r}$ and variance $\sigma^2$. i.e

$$r \sim \mathrm{N}\left(\hat{r}, \sigma^2\right), \tag{2.41}$$

then we can calculate $p$. Figure 2.9 shows how the probability changes for increasing values of $\sigma^2$ with different values of $\hat{r}$ (with $d = 1$). The shape of the curves in the figure matches Balch et al.'s Figure 1, even though dynamical simplifications have been made. The Figure shows that as $\sigma^2 \to \infty$, $p \to 0$ even when, deterministically, the objects would collide. Hence, as the uncertainty of the estimate increases, the probability of collision decreases, and we can be more confident that there will not be a collision.

This result is illogical as it is the equivalent of the proverbial ostrich becoming, statistically, more confident if it is safely sticking its head in the sand, meaning its uncertainty about the position of the predator was at its maximum, than if it found out exactly where the lion is and running away. This problem is known as the *False Confidence Theorem* First explored by Balch et al. [58], but also by Martin [162], and Carmichael and Williams [163], the false confidence theorem refers to the problem that, due to probability dilution, it is possible to become falsely confident about an event occurring even though the uncertainty about the prediction is large. Principally it is a result of using objects of aleatory uncertainty to model epistemic uncertainty. We can attempt to fix this problem by using PBA objects

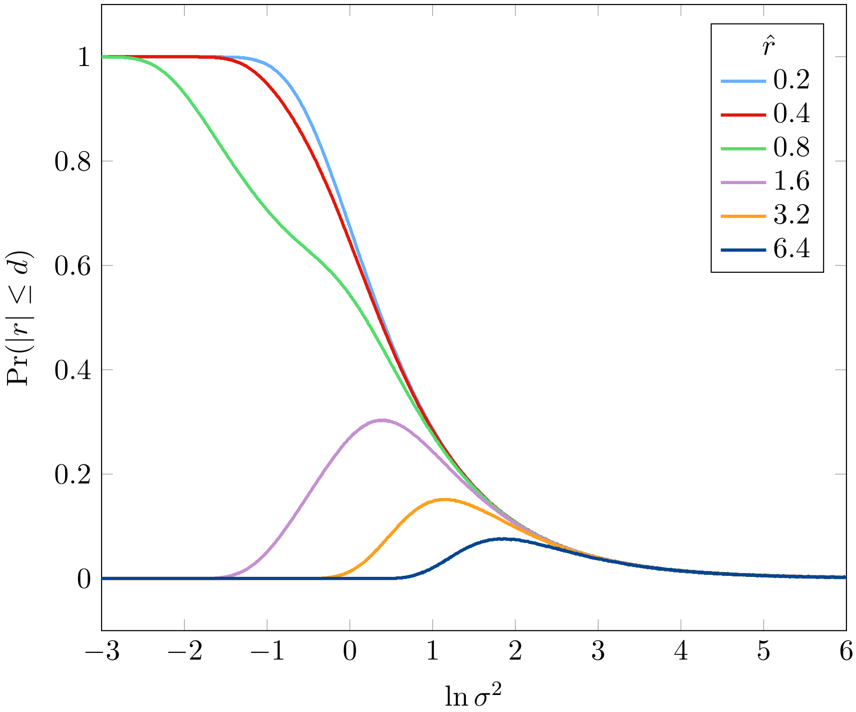


Figure 2.9: How the probability of collision, $p$, for two objects of diameter $d = 1$ changes with increasing $\sigma$ for different $\hat{r}$.

better suited to characterising epistemic uncertainty.

### An Interval Fix?

The most natural way of characterising epistemic uncertainty is by using intervals. For instance, we may only know the initial position of a particle with some error in each coordinate, $\mathbf{x} = (x_0 \pm \epsilon_x, y_0 \pm \epsilon_y)$, or there could also be uncertainty in the velocity of the particle $\mathbf{x}'(t)$.

There are then three different scenarios:

(i) We can be confident that the particle will miss the target when the entire final interval is missing the target, such as particle A in Figure 2.10.

(ii) We can be confident that the particle will hit the target. This can be concluded when the entire final interval is within the target, such as particle B in Figure 2.10.

(iii) We cannot have confidence about whether the particle will hit or miss the target. If some of the interval is hitting the target and some is missing, such as particle C in Figure 2.10, in which case we can only say the particle *might* hit the target, the

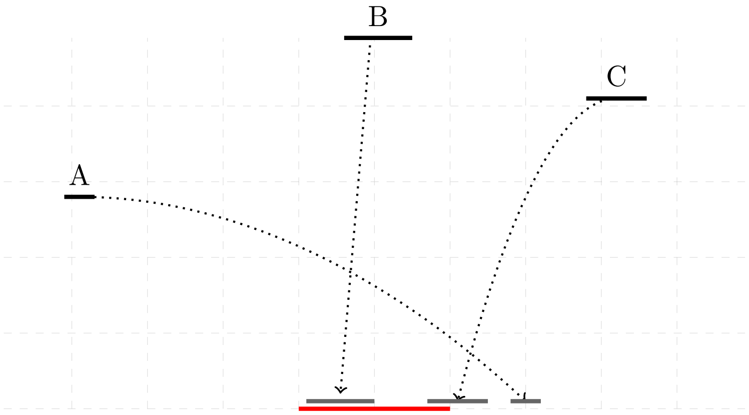


Figure 2.10: We can say with certainty that A will miss and that B will hit, whereas we can only say that particle C *might* hit the target

probability of the particle hitting the target is $p = [0, 1]$.

In high uncertainty cases, it may be the case that we are just unable to say whether it will hit or not. This uncertainty may seem undesirable, but it is the logical result and is natural of the admission of the epistemic uncertainty in the dynamical parameters. If we do not know the position or trajectory of the particle with sufficient precision, then we *should* not be able to infer whether it will hit the target.

### A P-Box Fix?

We can also reconsider the example when, instead of using a probability distribution, we use a probability box. A p-box is a better characterisation for the epistemic uncertainty in the measurement, as these structures can characterise both the epistemic and aleatory uncertainties that are present [140].

We can reconsider our estimate for the position as a p-box bounded by both the minimum and maximum variance cases,

$$r = \mathrm{N}\left(\hat{r}, \left[0, \sigma^2\right]\right). \tag{2.42}$$

Such a change means that instead of $p \to 0$ if $\sigma \to \infty$, there are two different limits of the uncertainty,

$$\lim_{\sigma \to \infty} p = \begin{cases} [0, 1] & \text{if } |\hat{r}| \leq d \\ \left[0, \frac{1}{2}\right] & \text{if } |\hat{r}| > d. \end{cases} \tag{2.43}$$

These differing limits provide a different interpretation of whether the particle will hit the target. When faced with considerable uncertainty, one would never be able to conclude that

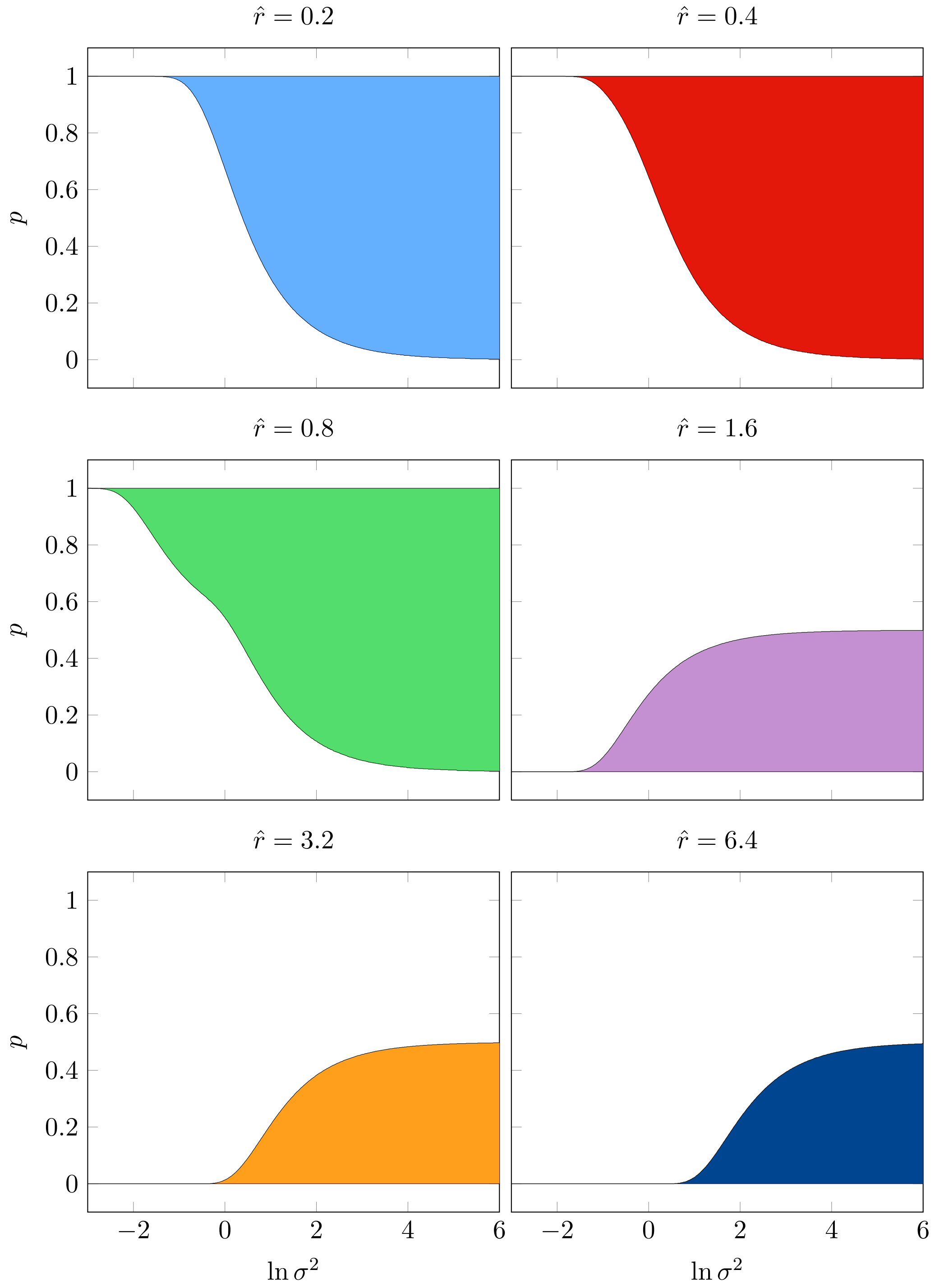


Figure 2.11: How the probability of collision, $p$, for two objects of diameter $d = 1$ changes with increasing $\sigma$ for different distances of closest approach, $r(T)$, using the imprecise probability method.

there would be no collision. This is an appropriate and sinsible conclusion based on the assumtions made, unlike those implied by Figure 2.9.

### Discussion

This example shows the importance of distinguishing the two different types of uncertainty. In the first case, the analysis quantifies the epistemic uncertainty through probability distributions. In effect, this treats the uncertainty as though it is aleatory uncertainty and leads to the illogical result that as the uncertainty increases, the probability of the particle hitting the target decreases. The problem is a result of treating epistemically uncertain values as though they are aleatorically uncertain and using probability distributions, in a frequentist sense, to model the uncertainty. This leads to illogical results as demonstrated in the example above, as well as some of the examples in Chapter 1. Using objects better suited to characterising epistemic uncertainty, such as intervals or p-boxes, alleviates these issues. Making sure that the uncertainties are used correctly within calculations will be important in making algorithms more humane, increasing the accessability of the algorithm and applicability of the numbers used.

# Chapter 3

# Towards an Automatic Uncertainty Compiler

***Preface*** A version of this text is currently due to be published by the *International Journal of Approximate Reasoning.*

A pre-print version of this chapter is available as Nicholas Gray, Marco De Angelis, and Scott Ferson. The Creation of Puffin, the Automatic Uncertainty Compiler. *arXiv:2110.10153 [cs, stat]*, October 2021. http://arxiv.org/abs/2110.10153.

Part of this work has also been publised as Nick Gray, Marco De Angelis, and Scott Ferson. Computing With Uncertainty: Introducing Puffin the Automatic Uncertainty Compiler. In V. Papadopoulos M. Papadrakakis, G. Stefanou, editor, *3rd ECCOMAS Thematic Conference on Uncertainty Quantification in Computational Sciences and Engineering*, pages 487–497, Heraklion, Greece, 2019. doi: 10.7712/120219.6354.18702.

As discussed in Chapter 1, modern science and engineering are all about numerical calculation, and increaingly work is being done taking advantage of the power of computers. The algorithms that are used often rely on precise, abstract numbers – without any associated uncertainty and units. Scientists and engineers need to make calculations even when there is uncertainty about the quantities involved, yet the tools they commonly use do not allow this to be done intrusively. As a result, many analysts work with computer codes that do not fully account for uncertainties.

Because analysts are typically unwilling to rewrite their codes – due to the time constraints or a lack of understanding of the uncertainty calculi available – various simple strategies have been used to remedy the problem, such as elaborate sensitivity studies or wrapping the programme in a Monte Carlo loop. This alows *non-intrusive uncertainty propagation*. These approaches treat the programme like a black box because users consider it uneditable. However, whenever it is possible to look inside the source code, it is better characterised as a crystal box because the operations involved are clear but fixed and unchangeable in the mind of the current user. An alternative approach is *intrusive uncertainty propagation* where the uncertain variables are replaced with objects that characterise their uncertainty (such as those described within Chapter 2) and have calculations directly performed on them. Other changes to the code may be required to allow for optimal performance.

An uncertainty compiler is a tool that automatically translates original computer source code lacking explicit uncertainty analysis into code containing appropriate uncertainty representations and uncertainty propagation algorithms. It handles the specifications of input uncertainties, and inserts calls to intrusive uncertainty quantification algorithms in the library. The uncertainty compiler can apply intrusive uncertainty propagation methods to codes or parts of codes and, therefore, more comprehensively and flexibly address epistemic and aleatory uncertainties.

## 3.1 Uncertainty Analysis: A Roll of the Dice?

The most common approach to quantify uncertainty is to wrap code within a Monte Carlo (MC) shell. In this approach, the calculations are repeated with random values for selected input variables. This is done for many iterations, and the distribution of resulting outputs can be analysed. Such tools exists in many programming languages: DAKOTA for C++ [164], COSSAN [165] and UQLab [166] for MATLAB or UQpy for Python [167]. Olivier et al. [167] give an excellent overview of many more software packages available for non-intrusive uncertainty quantification.

Several pitfalls make MC analyses unsuitable for uncertainty propagation, especially for epistemic uncertainty [168, 169]. To highlight these, we shall consider an example.

Suppose we have five variables, $x_1, x_2, \dots x_5$, which are known to all have a value between 0 and 1, but no further information is known about the values. Suppose we need to perform the calculation,

$$y = x_1 + x_2 + x_3 + x_4 + x_5, \tag{3.1}$$

with the knowledge that some bad thing will happen if $y \geq 4.5$. Numbers can be randomly generated for $x_1, x_2, \dots,$ and these can be used in order to calculate the value of $y$ for $N$ iterations. After this is complete, we can plot a histogram to show the distribution for $y$. Since we do not have any information about the distribution for $x_1, x_2, \dots x_5$, it seems sensible to assume that all values are equally likely and use a uniform distribution. Figure 3.1 shows these histograms for various $N$. From this, we can see that as $N \to \infty$, the histogram resembles a normal distribution.

Whatever the number of replications used in the simulation, we can estimate the probability of the bad thing happening. With $10^6$ replications, this estimate is $\Pr(y \geq 4.5) = 2.53 \times 10^{-4}$. However, it seems reasonable to consider whether we have confidence that the event is so rare. We had no information about the distributions of the five values except that they were between 0 and 1. Nor did we know what dependencies there might be between the variables. From this information, we cannot rule out the possibility that each $x$ value is much more likely to be closer to 1 than 0. Nor can we exclude that there is some dependence between the $x$ values implying that if $x_1$ is high, then all the others are also likely to be high.

Several engineering failures were due in part to underestimating risks in ways similar to this example [170, 171]. Before the 1986 Challenger Disaster, NASA management had predicted the probability of failure with loss of vehicle and crew as 1 in $10^5$ flights [172]. This turned out to be a gross underestimation of the actual risk, which after the fleet's retirement stood at 2 in 135. The Fukushima Daiichi nuclear disaster was due in part to underestimating the risk of a tsunami of the magnitude that caused the disaster and failing to understand that collocating the backup generators created dependence that destroyed the planned engineered redundancy when the site was flooded during the event [173, p. 48].

The MC method presented in Figure 3.1 is a simplistic approaches. There are many different approach that could have been taken to use MC to propagate the uncertainty through Equation 3.1 that are more advanced. It is also possible to use copulas to introduce dependencies between the random samples [174, 175]. There are several different sampling techniques that could be used to generate the MC samples and offer analytical stratergies for computing distributions

Latin Hypercube Sampling (LHS) is a popular method which aims to be representative of the true variability of the inputted numbers [176–178]. LHS is performed by dividing the range of each random variable into $N$ bins with equal probability mass, where $N$ is the required number of samples, generating one sample per bin, and then randomly pairing the samples. The samples generated are uniformly distributed over each marginal distribution [167].

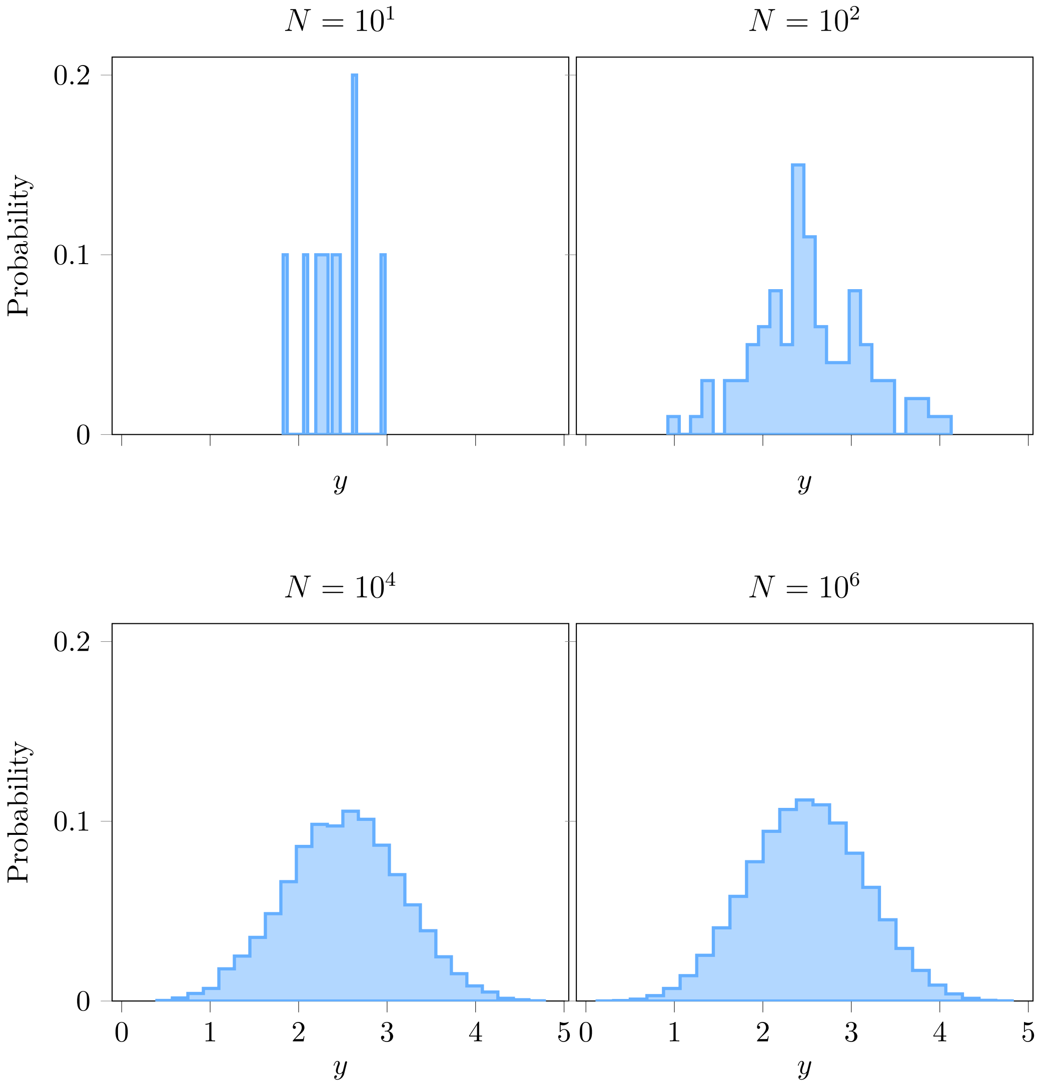


Figure 3.1: Normalised histogram for the Monte Carlo simulation of Equation 3.1 for increasing number of iterations.

There are also stratification methods that aim to divide the parameter space into a set of disjoint and space-filling strata that draw samples from the strata. The aim is to improve the space filling properties of the sampling and the method allows for weighted samples [179]. Refined stratification adaptively updates the stratification by dividing the existing strata in order to improve convergance [179].

Aside from MC there are other probabilistic approaches that could be used for uncertainty propagation. For instance, adaptive stochastic collocation techniques may have been better at finding the extreme values in the above example [180]. Or polynomial chaos expansion could have been used, although this technique would have required modifying the original source code [181].

Depending on the needs of the analyst, it may be the case that MC simulations may be suitable to solve their uncertainty quantification problems. After all, MC techniques are easy, efficient, and give insights into the aleatory uncertainty that may be present within simulations [182]. However, if there is epistemic uncertainty, since all of these different approaches ultimately rely on treating the all the uncertainties in a probabilistic way [132, 169] and false confidence may occur (see Section 2.6.2 or [58]).

Performing uncertainty analysis by simply wrapping a simulation code in a MC loop may not give a complete account of the uncertainties present within a simulation. The probabilities of extreme events are difficult to correctly estimate when there is no information about the distributions of input variables or any inter-variable dependencies.

An alternative to MC is intrusive uncertainty analysis, where the uncertainty is included directly within the code and calculations. There are objects that could be used within an automatic uncertainty compiler, including, but not limited to: probability bound analysis (as discussed in Chapter 2), second-order distributions or meta-distributions [146, 147], fuzzy numbers [148], possibility distributions [149], consonant structures [150], info-gap models [151]. All of which have numerous applications within engineering [132]. However, given the aim of this Chapter is to dicuss the feasibility of an automatic uncertainty compiler we shall limit this discussion to PBA.

## 3.2 An Automatic Uncertainty Compiler?

Strategies are needed that automatically translate source code into code with appropriate uncertainty representations and propagation algorithms. Perez et al. [183] introduced a MATLAB toolbox to perform automatic uncertainty propagation based upon the unscented transform. However, more general approaches are needed. This section discusses the prospect of creating an automatic uncertainty compiler to perform intrusive uncertainty analysis with crystal box codes (when the source code can be viewed but is considered fixed and uneditable). Such a compiler would handle the specifications of input uncertainties and insert calls to an object-oriented library of intrusive uncertainty quantification (UQ) algorithms. This approach could theoretically work with any computer language and any flavour of intrusive uncertainty propagation.

Such a compiler will have to perform several tasks:

1. Read the input source code to identify the variables in any assignment operations which may have uncertainty associated with them.
2. Replace or modify some or all of these assignments according to options and specifications provided by the user,
3. Translate the expression trees, with amended assignments, into the target language equipped with its intrusive UQ library, and
4. Optimise the output code to ensure that code will run robustly and not be artifactually inflated.
5. Once this trans-compilation has occurred, the output script can be run identically to the original code. As the uncertainty quantification is included directly within the simulation, the outputs will contain expressions of the associated uncertainty and be guaranteed to include the correct answer.

Further details of the precise requirements for each of these steps will follow, but, at this stage, it is helpful to consider a simple pseudocode script, as shown in the top left corner of Figure 3.2. In this plot, the first step (represented by the blue line) requires the compiler to read the input script and then detect and extract the assignment operations. These include lines 1 and 2, but not those that assign a value based upon a mathematical expression (line 5), a function or directly from another variable. In theory, such variables could also be edited by the user.

These extracted variables then need to have any required uncertainty added. These uncertainties must then be translated to the source language and merged with the original script to produce a new script, as shown at the bottom of Figure 3.2. This translation may include altering any functions that depend on the amended variables. In this case, the infix operators in the definition of `d` in line 5 (`+,*`) have been recognised and replaced with an explicit call to the UQ library functions (`add`, `mul`) which also have as an argument the dependence operation that is to be used. The lower panel of Figure 3.2, the value ‘f’ of this argument corresponds to making no assumption about the intervariable dependence between `a` and `b`, and perfect dependence (comonotonicity) between their product and the variable `c`.

## 3.3 Reading and Modifying

The first stage that the compiler will need to perform is to be able to read the script and detect the assignment operators that might have some uncertainty about them. Only numeric objects should be highlighted, not characters, strings or other non-numeric classes. In strongly typed programming languages like C, FORTRAN, and Pascal, the task of distinguishing numeric from other types of objects is easy. In Python, R, Julia or any other dynamic duck typed language, the object type is not detectable until runtime and can even change during execution. This means that it may be difficult for the reader to detect*.

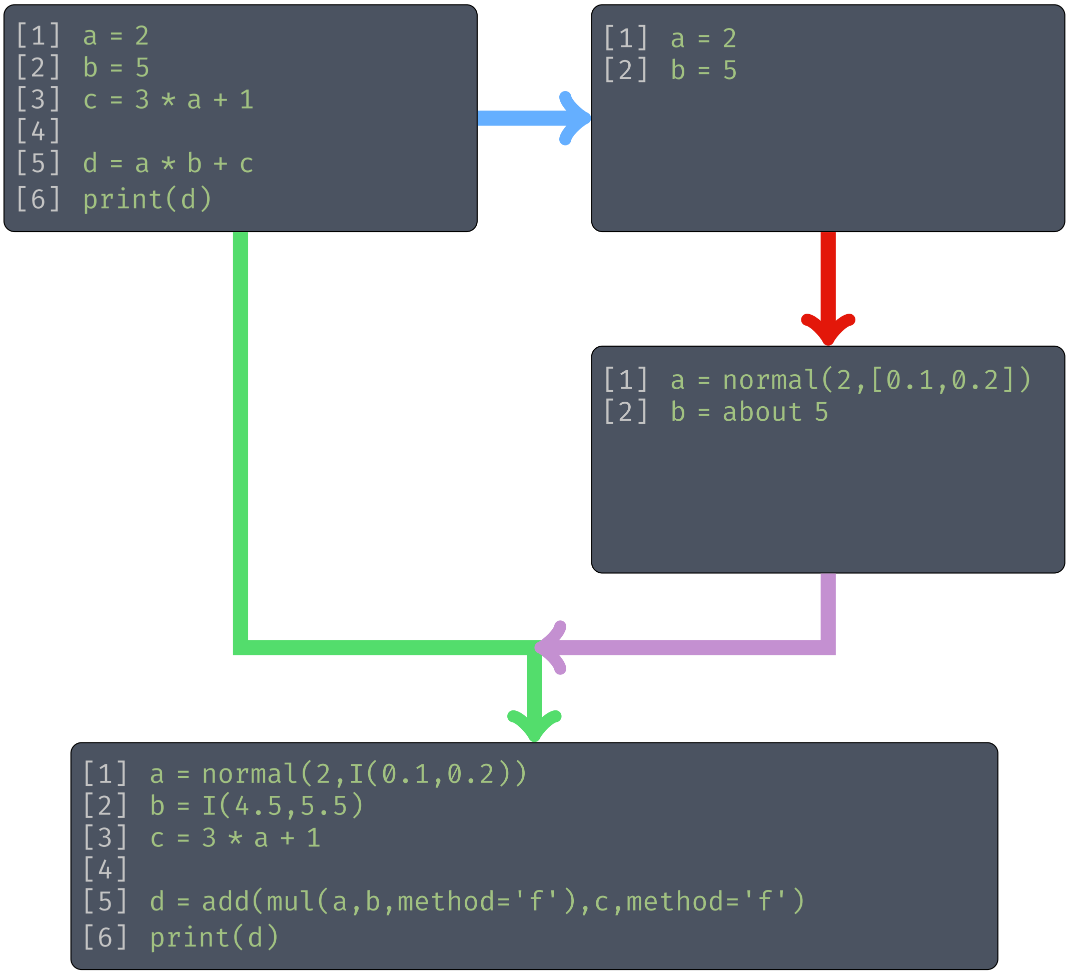
Objects that are collections of numeric values, such as arrays, lists and dictionaries, will require special attention. It may be necessary to look at what is inside the lists and highlight those with numeric objects within to allow the individual objects to have any uncertainties added. Alternatively, it could be the case that the list represents a vector within which all elements have the same uncertainty, something which should be possible.

Once these objects have been extracted, they should be expressed within a language that enables users to specify the uncertainties associated with the extracted variables. This language should be simple and independent of the source language and does not need to be a textual programming language; instead, it could be a visual language as part of a graphical user interface. Every type of uncertainty that is expressible within this language must be supported by an object constructor in the uncertainty library written in the source language.

Additionally, if there is any dependency between any two variables, then the language needs a way for users to specify what these are. A natural way to do this is for the reader to produce a matrix of all linked variables and their associated dependencies. For example, as line 3 of Figure 3.2 specified that `c = 3 * a` the compiler could detect that there is perfect dependence between them. The user can then edit this matrix to specify any known dependencies, as shown in Table 3.1. Fréchet would be used for calculations between variables that do not have specified or detectable dependencies.

In general, dependence between uncertain quantities can be expressed through the use of correlation coefficients or copulas or bounds on copulas more generally [175, 184–186]. This can include named copulas such as independence, opposite and perfect, as discussed in Sec-

*However, when it comes to translating and writing duck typed languages are easier to implement.



```
a = 2
b = 5
c = 3 * a + 1

d = a * b + c
print(d)
```

```
a = 2
b = 5
```

```
a = normal(2,[0.1,0.2])
b = about 5
```

```
a = normal(2,I(0.1,0.2))
b = I(4.5,5.5)
c = 3 * a + 1

d = add(mul(a,b,method='f'),c,method='f')
print(d)
```

Figure 3.2: The steps that an uncertainty compiler needs to take in order to take the simple pseudocode script in the top left, extract the assignments (blue line), add in the user-specified uncertainty (red line), translate the uncertainties into the original language (pink line) and then merge into the original script (green line), making appropriate additional changes so that the code runs optimally.

tion 2.3 and shown in Figure 2.2, or other copula families parameterised by a numerical correlation coefficient. Independence implies the correlation is zero, although zero correlation does not imply independence. Likewise, a correlation of one implies perfect dependence, but, depending on the copula family, perfect dependence may not imply correlation one. The symbol ≡ can be used to indicate that the variables are equal in value, i.e., equal in distribution and perfectly positively correlated.

| | a | b | c | d | e |
|---|---|---|---|---|---|
| a | | i | ≡ | f | o |
| b | i | | i | f | 0 |
| c | ≡ | i | | f | o |
| d | f | f | f | | f |
| e | o | 0 | o | f | |

Table 3.1: Matrix showing dependencies between several variables. (f - Fréchet, i - Independence, o - Opposite, ≡ - Equal in value, 0 - implies that there is no correlation between the two variables.)

The dependency matrix can be checked for feasibility by asserting that it is positive semi-definite and that there are no conflicting dependencies within the table. For example, the matrix shown in Table 3.2 is not logically consistent for continuous variables. This is because a high value of x implies a high value of z, since they are positively dependent on each other. Meanwhile, a high value of z implies the value of y must be low due to their opposite dependence. Hence, there must also be dependence between x and y, not the independence specified in the table.

| | x | y | z |
|---|---|---|---|
| x | | i | p |
| y | i | | o |
| z | p | o | |

Table 3.2: Dependency matrix that does not make logical sense. (i - Independence, p - Perfect, o - Opposite)

In order to make specifying the uncertainty as simple as possible, users should be able to input their uncertainties using natural language expressions such as *about* or *almost*. As discussed in Chapter 1, Sections 1.1 and 1.3, humans are more likely to express their uncertainties in terms of hedged expressions around a round number rather than as a percentage or probability. Table 3.3 lists some hedge words and their possible interpretations. Hedge words can be interpreted as intervals or p-boxes [26, 187].

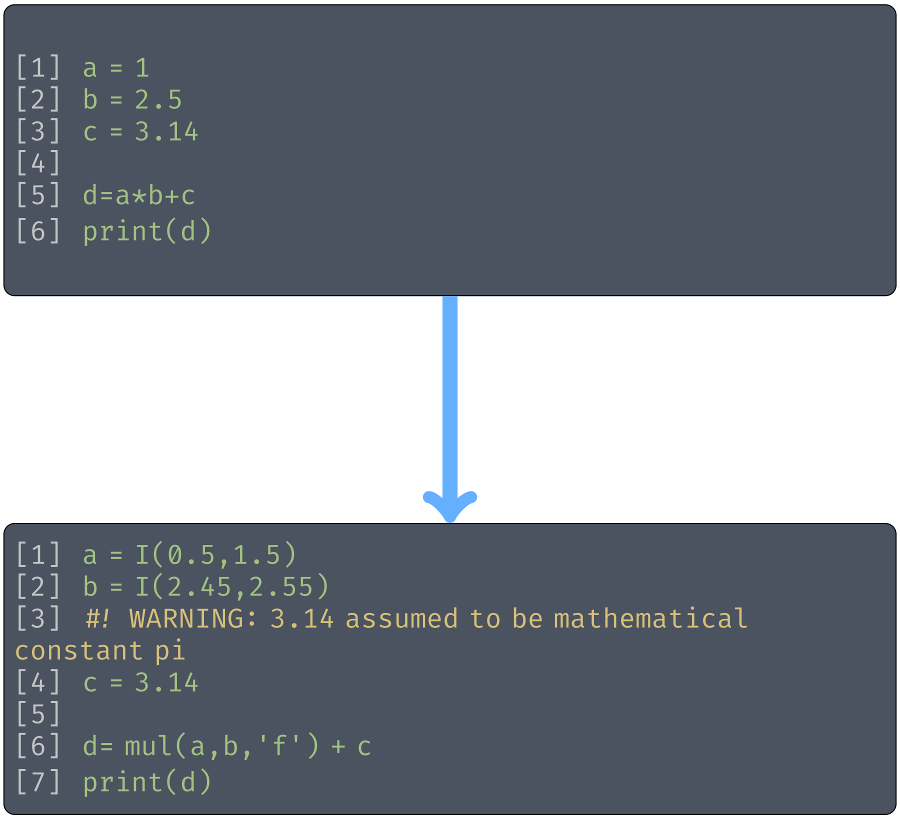
Often it is the case that all code is not contained within one script. For example, classes and functions are often placed in other files to improve readability or avoid repetitions.

| Hedged Numerical Expression | Possible Interpretation |
|---|---|
| ABOUT($x$) | $[x \pm 2 \times 10^{-d}]$ |
| AROUND($x$) | $[x \pm 10 \times 10^{-d}]$ |
| COUNT($x$) | $[x \pm \sqrt{x}]$ |
| ALMOST($x$) | $[x - 0.5 \times 10^{-d}, x]$ |
| OVER($x$) | $[x, x + 0.5 \times 10^{-d}]$ |
| ABOVE($x$) | $[x, x + 2 \times 10^{-d}]$ |
| BELOW($x$) | $[x - 2 \times 10^{-d}, x]$ |
| AT MOST($x$) | $[0, x]$ |
| AT LEAST($x$) | $[x, \infty]$ |
| ORDER($x$) | $[x/2, 5x]$ |
| BETWEEN $x$ AND $y$ | $[x, y]$ |
| $k$-out-of-$n$ | $\mathrm{KN}(k, n)$ |

Table 3.3: Hedge expressions and their mathematical equivalent. Note: $d$ is the number of significant figures of $x$.

Ideally, the compiler would be able to read several scripts simultaneously and allow users to express the uncertainty whilst remembering the context for all the individual objects. It is also often the case that scripts read data from other files when running. Under this scenario, it would be difficult to express the uncertainty directly within the script, although import functions could be modified to add in the uncertainties. For instance, anytime a floating-point number is read from the file, its significant digits could be interpreted to specify an interval around the value. So, for example, the value '3.56' would be understood as the interval [3.555, 3.565]. Another approach might be to parse the data file and add the uncertainties into the file directly. Doing so would require changing the import function to be able to handle uncertain data files.

It is also possible to completely automate the uncertainty compilation. Under default settings, this would replace floating-point constants with intervals interpreted from the significant figures used in the source code assignments and use that information as a proxy for the uncertainty (for an example, see Figure 3.3). In this mode, each individual task happens concurrently without requiring any further input from an end-user. When using this mode, the compiler must tread carefully around mathematical constants such as $\pi$ or $e$ for which there is no uncertainty. In the example, the compiler has printed a warning above `c = 3.14` to highlight the assumption that has been made. The compiler, it would allow users to specify what values are precise constants. It is also worth noting that since `c` has no uncertainty the addition operation does not need to be replaced with a function with specified dependence arithmetic and Fréchet has been used for the dependence between `a` and `b`.



```
a = 1
b = 2.5
c = 3.14

d=a*b+c
print(d)
```

```
a = I(0.5,1.5)
b = I(2.45,2.55)
#! WARNING: 3.14 assumed to be mathematical constant pi
c = 3.14

d= mul(a,b,'f') + c
print(d)
```

Figure 3.3: Result of using the automatic uncertainty compiler on a simple pseudocode script.

For-loops and functions are potential stumbling blocks for any uncertainty compiler. Figure 3.4 shows a simple pseudocode script with a function and a for-loop. Within fir first for-loop, each `i` is simply a control variable with a start and end variable. The individual value of `i` is irrelevant and, as such, would have no uncertainty about it. The second for-loop is a foreach-loop, implying that the code needs to do something for each value within some iterable object. Under this scenario, it may be the case that there is uncertainty about the object within the list. In Figure 3.4, the code is setting the value of `initial_velocity` as each value within the list for each iteration of the for-loop. It may be the case that there is uncertainty within the object, ergo the compiler should recognise this and allow users to change the code such that the objects within the list can have uncertainty added to them.

Local variables within functions will also have to be handled. For example, the function in Figure 3.4 has two local variables, `s` for the distance that the object travels and `g` for the acceleration due to gravity. It is conceivable that both of these variables have some uncertainty associated with them, and the compiler needs to be able to detect the variables and offer the ability to edit them so that uncertainty is handled.

```
[ 1] def calculateVelocity(u):
[ 2]    s = 10
[ 2]    g = 9.81
[ 3]    v = sqrt(u^2 + 2*g*s)
[ 4]    return v
[ 5]
[ 6] x = 1
[ 7] for i in (1:10):
[ 8]    x = x + 1
[ 9]
[10] for initial_velocity in (0.3,0.4,1.1,2.3,3.7):
[11]    final_velocity = calculateVelocity(initial_velocity)
[12]    print(final_velocity)
[13]
[14] a = 1
[15] b = 0.5
[16] if a < b:
[17]    c = 1
[18] else:
[19]    c = 2
[20]
```

Figure 3.4: Pseudocode script with functions, if-statements and for-loops.

If-statements and other logical control structures may also pose issues for such an uncertainty compiler. In line 16 of Figure 3.4, the logical operators within the statement would need to change to ensure that the statement runs as expected (see Sections 2.1 and 2.2).

Ideally, the analyst would decide what should happen if the statement `a < b` returned a [0,1] result by using the adverb operators discussed in Section 2.1. This may require additional editing to deal with situations where an uncertain result should be handled differently from a certain true or false or when p-boxes return probabilities for the comparison.

## 3.4 Translating and Optimising

Once the uncertainty has been specified, it will need to be translated into the target language. Assuming that there exist appropriate uncertainty objects within the source language, this need not be difficult. The more difficult task is modifying the code such that it runs optimally and the amount of artifactual uncertainty is reduced. For instance, amendments to the code will be needed to ensure that dependencies are handled correctly. There are a couple of approaches that could be used.

The simplest approach is to replace the infix operators with the function calls that take as an argument the dependency calculi needed for the operation. For instance, in Figure 3.2, the calculation

```
d = a * b + c
```

has been replaced by the compiler to

```
d = add(mul(a, b, method = 'f'), c, method = 'f')
```

where the `add` and `mul` contain keyword arguments that tell the code what arithmetic to use. In this approach care has to be taken to make sure that precedence, left recursion and associativity of the operators is handled correctly [188, pp.69–72].

Any dependence between `a` and `b` would have to be initialised. This could be done by adding a function after both `a` and `b` have been assigned or by modifying the assignment for `b`. In this case, as no dependence is known to the compiler, Fréchet has been used.

**Smart dependency tracking** – where the varible themselves workout how to handle the uncertainty arithmetic with other uncertain objects at runtime – would be the zenith when wrangling the issue. Whilst initial dependencies would still have to be specified, under smart dependency tracking, no other changes would need to be made to the code by the user or compiler for the dependencies to be handled optimally. In the output script of Figure 3.2, line 5 would not need to be changed as the variables would remember which other variables they depend on and, when the calculations are performed, know what dependency calculus to use. This knowledge would have the be included within the class definitions for the uncertain objects and requires all objects to store the ancestry of themselves, knowing the dependence between themselves and their ancestors and be able to access the dependencies between the initial variables. This approach would have demands on memory and computational time. The result of `a * b` would know that it depends on `a` and `b`. When `c` is added to the product, it would be detected that `a * b` and `c` both have `a` as a common ancestor and therefore use the appropriate addition methodology would be ascertained and used.

Another major consideration for a uncertainty compiler will be mitigating against the repeated variable problem (see Section 2.3). The compiler will need to be able to rearrange equations such that repeated variables are removed whenever possible. The most basic approach will require having a directory of multiuse to single-use expressions. Another smarter approach would be to have a symbolic algebra system that can rearrange to a single-use expression on the fly. The simplest version of this is for it to happen just across one line, for instance, replacing

$$d = ab + ac \tag{3.2}$$

with

$$d = a(b + c). \tag{3.3}$$

These rearrangements are often not obvious and may require algebraic tricks, i.e.

$$w = \frac{x + y}{1 - xy} \tag{3.4}$$

can be rearranged into the single-use expression

$$d = \tan\left(\arctan(x) + \arctan(y)\right). \tag{3.5}$$

This problem is made more difficult if there are repetitions across multiple lines. In Figure 3.2, `c = 3 * a + 1` is used in the calculation `d = a * b + c`. This equation could be rewritten as

```
d = a * b + 3 * a + 1
```

and the reduced to the single use equation

```
d = a * (b + 3) + 1.
```

Care would need to be taken to ensure that the right variable gets the rearrangement. Take the following kinematics equation to find the position $s$ of a particle at time $t$,

$$s = ut + \frac{1}{2}at^2, \tag{3.6}$$

where $u$ is the initial velocity and $a$ is the acceleration of the particle. This equation has a single repetition for both $u$ and $a$ but $t$ is repeated. If there is uncertainty associated with $t$ then this equation can be rearranged into a single-use expression

$$s = \left(\sqrt{\frac{a}{2}}t + \frac{u}{\sqrt{2a}}\right)^2 - \frac{u^2}{2a}. \tag{3.7}$$

This equation contains repetitions of $a$ and $u$ and, as such, may only be preferred if there is no uncertainty associated with either $a$ or $u$. If there is uncertainty associated with either, then it may be best not to perform the rearrangement or to intersect possible rearrangements to obtain the best possible expression.

There are additional issues that the compiler could arise when rearranging equations. For example in Equation 3.7, if the uncertainty about $a$ includes negative numbers, then $\sqrt{2a}$ is likely to be problematic. Additionally, there could be problems if $a$ straddles 0 because this would result in a division by zero. A strategy for dealing with this may be to perform the calculations using both Equation 3.6 and 3.7 and intersect them to remove any artifactual uncertainty.

Many computer languages that are not purely functional support functions that specify their parameters with “call by reference” or “call by object reference”, meaning the memory location of a value is passed to the function rather than a copy of the actual value. This convention can allow the function to change the values of those parameters in the calling routine, not just locally within the function. For example, Python passes the object reference to functions for all non-primitive objects, such as lists, data frames and NumPy arrays. Primitive objects in Python are: integers, floating-point numbers, Boolean, and strings. Handling languages that use the call-by-reference method of passing arguments will require care when translating.

In Figure 3.5, the simple python function takes as an input a numerical value, adds 1 to it using the inplace `+=` operator and then returns the addition. In the case where `x` is a primitive object (an integer), the value passes to the function and returns the calculation without modifying the original `x` value.

If the code was translated using an uncertainty compiler then this may cause an issue. In *Translation #1*, since `x` is now a (non-primitive) interval object then the objects reference gets passed to the function. This means that it *may*[†] be the case that as the local variable `a` gets modified the global variable `x` also gets modified. This behaviour can be seen within

[†] In my PBA for Python library this behaviour is not observed. This is because PBA for Python’s Interval class’ addition method returns a new Interval instance rather than modifying the interval inplace.

the example, the function has returned the addition but `x` has also been changed because the object is not copied or cloned—its reference has been shared. This issue has been fixed in *Translation #2* by the compiler recognising that `a += 1` may pose a particular problem and replacing it with `a = a + 1` as this reassigns a new object to the variable rather than to the location it references[‡].

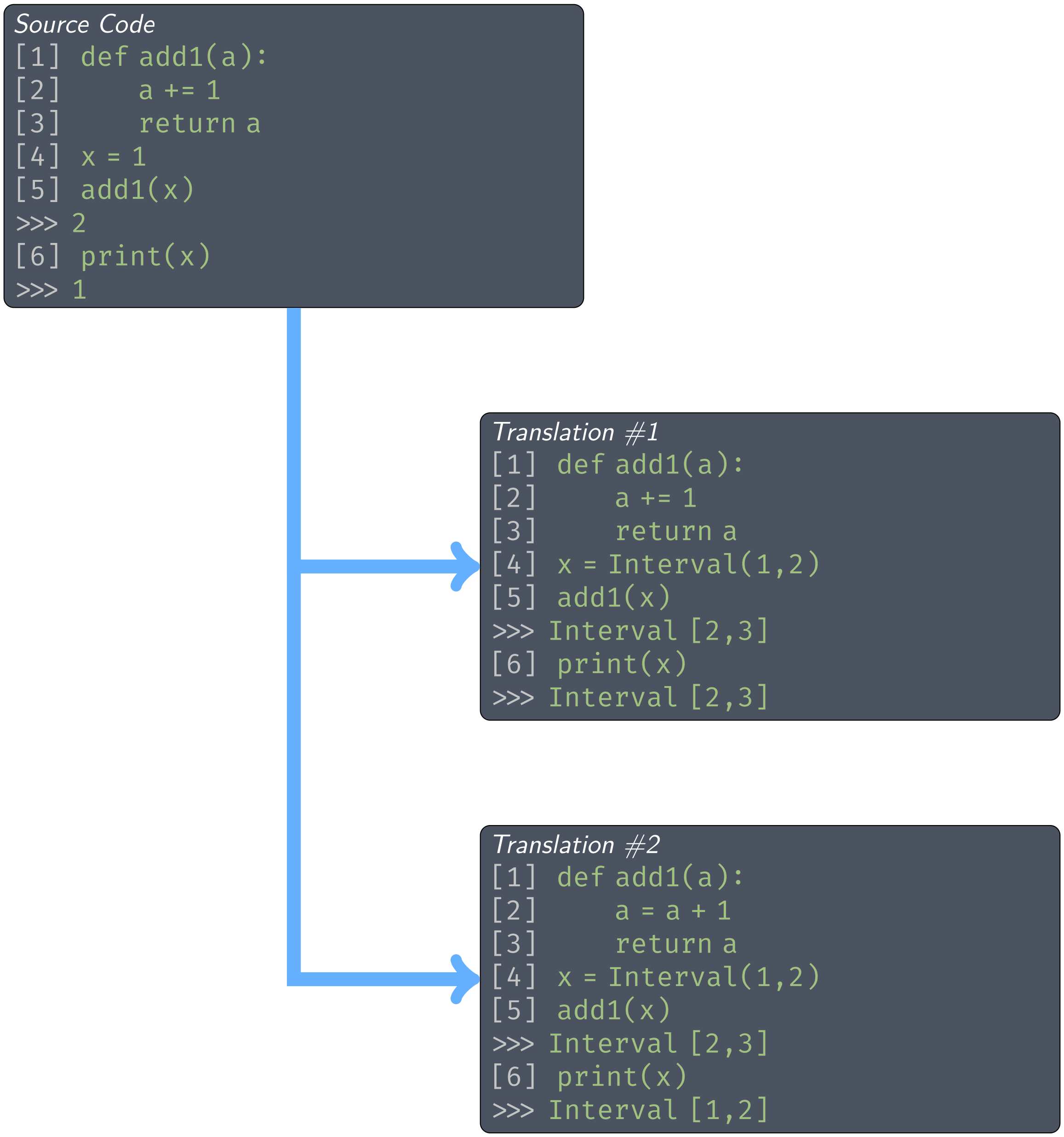
### 3.4.1 Hermeneutic Problems

Several problems could occur when translating a script because it can be difficult to understand a programmer's intent from the code. An example would be the assignment `c = a` as there are a several different interpretations as to what such a command implies when it comes to the uncertainty:

1. `a` and `c` are the same object but have been given different names for some reason. Under this scenario, they should be considered equivalent to each other, and therefore the calculation `a + c` could be rearranged to the single-use expression `2 * a`.
2. `c = a` could have been written as `c = 1 * a` and the `1` has been dropped as it would have had no mathematical impact on the calculation, this implies that they are perfectly dependent on each other in the same way that `c = -1 * a` implies negative dependence. Therefore, the calculation `a + c` would need to be performed using perfect dependence.
3. `c` is a copy of `a`; they have the same uncertainty, but their realisations are not necessarily related to each other whilst having the same distribution shape. This leads to two more distinct possibilities:
   (a) `a` and `c` are independent of each other and `a + c` should be convolved using independence.
   (b) No assumption is made about the dependence between `a` and `c` so `a + c` should be convolved using Fr'echet.

Knowledge of which of these scenarios is correct depends on the context of the script, something a compiler should not make an assumption about by itself.

[‡] It is worth noting that this may not always be desired behaviour. In Python `foo = foo + bar` is interpreted as `foo.__add__(bar)` whereas `foo += bar` is interpreted as `foo.__iadd__(bar)`. There may be difference between these two functions and it is not unimaginable the two functions have different outputs (although I struggle to imagine this impacting the numerical operations that an uncertainty compiler would be modifying). It is likely to be the case that this difference would impact the overhead of the calculation – which itself may have lead to the analyst prefering one form over the other (especially for NumPy objects).



*Source Code*

```
[1] def add1(a):
[2]     a += 1
[3]     return a
[4] x = 1
[5] add1(x)
>>> 2
[6] print(x)
>>> 1
```

*Translation #1*

```
[1] def add1(a):
[2]     a += 1
[3]     return a
[4] x = Interval(1,2)
[5] add1(x)
>>> Interval [2,3]
[6] print(x)
>>> Interval [2,3]
```

*Translation #2*

```
[1] def add1(a):
[2]     a = a + 1
[3]     return a
[4] x = Interval(1,2)
[5] add1(x)
>>> Interval [2,3]
[6] print(x)
>>> Interval [1,2]
```

Figure 3.5: Simple Python script modified by the uncertainty compiler. In Translation #1 the compiler has modified `a += 1` to `a = a + 1` to guard against errors caused by call by object reference.

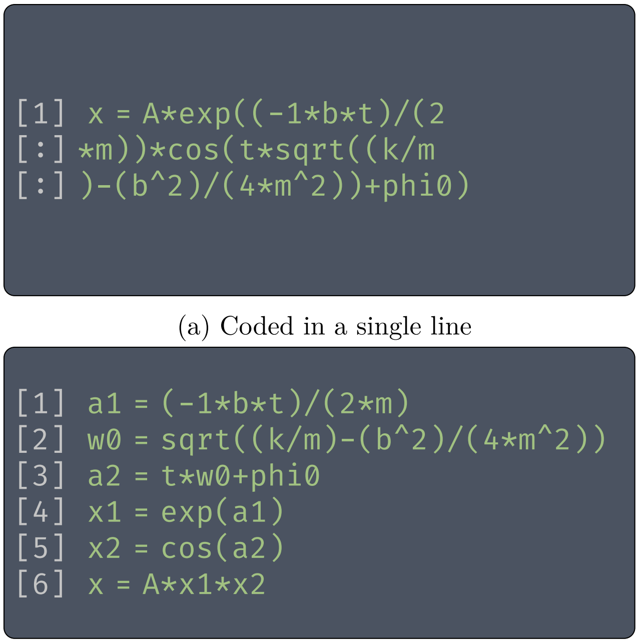
Another potential interpretation problem can occur because people naturally favour making their code readable when creating code. For example, the equation of motion for a damped harmonic oscillator can be given by

$$x''(t) + \frac{b}{m}x'(t) + \frac{k}{m}x(t) = 0 \tag{3.8}$$

where $b$ is a damping constant, $m$ is the mass of the oscillator, and $k$ is the spring constant. This equation can be solved analytically to find

$$x(t) = A \exp\left(\frac{-bt}{2m}\right) \cos\left(t\sqrt{\frac{k}{m} - \frac{b^2}{4m^2}} + \phi_0\right) \tag{3.9}$$

where $A$ is a constant and $\phi_0$ is the initial angle. In Figure 3.6 the equation has been coded in two different ways. In 3.6a the equation has been coded on a single line. As the equation is quite complicated, the programmer coding the equation would likely want to split it into multiple parts, as has been done in 3.6b. There are no mathematical differences between the two approaches as they will lead to the same value. Care would have to be taken breaking up equations in such a way that strategies would be needed to ensure that breaking the lines up would not have a detrimental effect on how the code operated. The dependency tracking throughout the split equation must also be assessed correctly. For example, if there were uncertainty about the damping constant $b$, it would be difficult to assess the dependence between lines 4 and 5, especially since the cosine function is not monotonic. It may be better to use other techniques to solve the ODE (Equation 3.8) such as VSPODE or VNODE [189–193]. Making such a change would again require knowledge of what the calculation is and what exactly it is doing, something unlikely to be obvious from parsing the code.



```
x = A*exp((-1*b*t)/(2*m))*cos(t*sqrt((k/m)-(b^2)/(4*m^2))+phi0)
```

(a) Coded in a single line

```
a1 = (-1*b*t)/(2*m)
w0 = sqrt((k/m)-(b^2)/(4*m^2))
a2 = t*w0+phi0
x1 = exp(a1)
x2 = cos(a2)
x = A*x1*x2
```

(b) Coded across multiple lines

Figure 3.6: Two different ways in which Equation 3.9 might be coded.

## 3.5 Writing and Running

Once the transcompilation has been performed, a new source code will have been created which can be compiled or run in the same way as the original script. Running this code may take longer and require more computational resources. However, it may still be beneficial in terms of computational cost over Monte Carlo as it only needs to be run once and not many times.

Once the code has been run, any outputs will themselves be uncertain. There are numerous ways in which this uncertainty can be expressed depending on the exact requirements specified by the analyst. Outputs could be the most likely value if this is sufficient for the analysis. Another simple approach would be to use the most appropriate hedged expression. For instance, it could return `about 3.2` if that is the best expression. This hedge would have to be the closest possible to the actual range without understating the uncertainty, whilst not being so large that it overstates the uncertainty.

For distributions, a simple approach would be to present the range or a 95% range (2.5th and 97.5th percentiles) of the output. This could go alongside other statistics about the distribution to present, for example, a five-number output that characterises the distribution (min, first quartile, median, third quartile, max). If the distribution family is known, it might be helpful to output the range, mean and variance.

Another way of characterising the distribution is to provide a verbose statement presenting a sentence or paragraph to describe the uncertainty. For instance, the output could be expressed as “a normal probability distribution p-box between -1.7 and 11 with a mean of 4.5 and a variance between 4 and 5” or “an uncertain number with unspecified shape between 0 and 10 with a median of at most 2.4, a mean of 1.2 and a variance of at most 10.6” Additionally, these statements could also come with a graphical representation of the output, such as shown in Chapter 2.

## 3.6 Conclusion

The idea of an automated uncertainty compiler is unlikely to be realised. Many codes are not simply single-line equations, and uncertainty propagation is not as simple as replacing every elementary operation with an interval or p-box equivalent. Numerous wrinkles have to be addressed, including:

- input specification (shapes and details),
- computational burden,
- repeated variables,
- dependencies,
- programmer intent, and
- suitable output.

It is also likely to be the case that no automatic uncertainty compiler would introduce perfect uncertainty translations. Manually editing code or creating a new uncertainty-aware script from scratch will likely outperform the automatic changes since, in general, hand-coding is always likely to outperform source-to-source translation [194]. It is, therefore, evident that any uncertainty compiler will likely be semi-automatic at best.

Perhapse the greatest challenge is that rearrangements will need to be made to equations to prevent artifactual inflation of the uncertainty. The compiler will also need strategies to handle inputs from other files, non-standard inputs, and code that spans multiple interconnected lines. The repeated variable problem and, more generally, incorrect handling of dependencies can artifactually inflate the uncertainty of complex computation. Even if the calculations are technically correct in the sense that they enclose the true uncertainty, the naive application of intrusive uncertainty quantification can sometimes yield results with massively inflated uncertainty that renders them practically useless. Having massive uncertainty is not the problem itself; the true uncertainty may actually be considerable. The problem is when the uncertainty artifactually depends on how the analysis was structured and does not reflect the features of the underlying computational problem.

An uncertainty compiler is perhaps the exact opposite of an optimising compiler, which aims to minimise a programmes execution time and memory requirements. Both of these will almost certainly increase by replacing objects with uncertain equivalents. For instance, if intervalising calculations increases the computational time five-fold,[§] a simulation limited by computational time may need to be scaled back. Distributions or p-boxes would be still more burdensome. Of course, efficiency is not always a critical issue, and this extra computational effort does pay for global uncertainty propagation and what computer scientists call automatic result verification [195]. Moreover, implementing modern uncertainty quantification techniques could be more efficient and comprehensive than simply embedding a deterministic computation inside a Monte Carlo shell with millions of replications.

Despite this, it is important to work towards the goal of an *automatic uncertainty compiler*. It is worth remembering that the analysts who will need one most may not be sure what normal distributions or intervals are and have probably never heard of a p-box. Nor do they have the time to sift through hundreds or thousands of lines of code to extract and manipulate the various parameters and make the changes required to enable the code handle the uncertainty optimally. For such analysts, a software tool that can hand-hold them through making the appropriate edits will be of enormous value, and even partial solutions will be beneficial.

The general problem of uncertainty analysis is hard, and it is difficult to create software that comprehensively solves all these problems. The development of an uncertainty compiler is likely to be difficult. However, many practical problems are more straightforward than the most general problem, and when this is the case, it will be extremely useful to use a tool which is able to handle uncertainty analysis intrusively within code.

[§] These numbers are for entirely illustrative purposes and should not prime readers' expectations for the increased computational burden PBA methods introduce.

# Chapter 4

# Difficulties in Quantifying the Uncertainty of Classification without Gold Standards

***Preface*** Material from N. Gray, D. Calleja, A. Wimbush, E. Miralles-Dolz, A. Gray, M. De Angelis, E. Derrer-Merk, B.U. Oparaji, V. Stepanov, L. Clearkin, and S. Ferson. Is "no test better than a bad test"? Impact of diagnostic uncertainty in mass testing on the spread of COVID-19. *PLoS ONE*, 15(10), 2020. doi: 10.1371/journal.pone.0240775 is used within Section 4.1.

Material from Nick Gray, Marco De Angelis, Dominic Calleja, and Scott Ferson. A Problem in the Bayesian Analysis of Data without Gold Standards. In *29th European Safety and Reliability Conference*, pages 2628–2634, Hanover, Germany, 2019 is used within Section 4.1 and 4.2.

Binary classification tests occur in many fields, from structural health in engineering to supervised learning in computer science or patient diagnosis in medicine. No tests are perfect, as they yield false positives and negatives. For instance, medical practitioners commonly diagnose a patient's health condition based on a diagnostic test, which in isolation is not definitive. The diagnostic result has some statistical uncertainty associated with detecting the true health state. Naively interpreting the result of a medical test can lead to an incorrect assessment of a patient's true health condition. The same is true for many binary classification systems. For the remainder of this chapter we shall refer to all binary classifcations as tests.

Tests can be evaluated by comparing the results of the test to the ground truth (the true classification). From this information, various statistics can be generated to assess the test's performance. However, in many instances, there is no gold standard test that can infallibly determine the ground truth. In medicine, there are many diseases for which there is no way to determine whether a patient has a particular disease conclusively. For instance, for patients with Giant Cell Arteritis, even after a biopsy has been undertaken, there is still uncertainty about the actual health state [196]. In machine learning, humans are often imagined as omnipotent oracles able to provide a correct classification, though they are usually imperfect [197–199].

In some situations gold standard information can only be gathered from some classes, yet for others it is unknown. For example, some prison authorities use classification algorithms to assess whether a prisoner is likely to reoffend when released from prison on parole (see Section 1.3). As authorities are unwilling to release prisoners if the test says they are likely to return to crime, it would be impossible to know whether the recidivism test was accurate. If a prisoner tested positive and thus remains imprisoned, there is no available data on whether they would have reoffended had they been released. Therefore, we could never know if it was a true positive or a false positive.

There is debate within the literature about whether the gold standard should be considered as the test that can accurately determine the true classification [200], or simply the best available test [201]. The latter definition is helpful in situations where the gold standard test changes over time as diagnostic technology improves. In medicine, there is also the question about whether it is valid for a disease not to have a gold standard [202]. What does it mean to test positive for a disease if there is no way to say definitively whether one has a disease? Is it necessarily the case that a positive gold standard result always implies some adverse pathological condition? [203] For the present analysis, we shall consider a gold standard to be the test that can identify the ground truth with perfect accuracy.

Numerous authors have made attempts to characterise binary classification tests without a gold standard; however, there is no widely accepted approach to dealing with this problem [204]. Rutjes et al. [205] conducted a literature review and found that these methods fall under four basic categories:

1. **Correct imperfect gold standard**. In this case, the gold standard is known to be imperfect, but no better test is available or accessible. In this instance the test

is better described as a pyrite standard. Adjustments are made to the test to estimate test statistics or perform a sensitivity analysis. These adjustments require prior information to be known about the performance of the gold standard test.

2. **Impute for missing data about the gold standard**. This is used when there is a true gold standard that is not used for whatever reason; the cost might be too high, the test too destructive to perform, or it is practically unobservable for some other reason. Imputation requires assuming that the missing values are *missing at random* or *missing completely at random* to be made and the fraction of the missing gold standard results to be small compared to the overall sample size.

3. **Construct gold standard**. Multiple none gold standard tests are combined to find the ground truth. This construction can be performed by expert elicitation, latent class analysis or other deterministic rules.

4. **Validate index test result**. The confusion matrix paradigm is abandoned, with the test results validated against other relevant characteristics, and different statistics are calculated, including event rates, relative risks and other correlation statistics. For example, in a medical context, the test may be validated by assessing whether it can predict who will benefit from an intervention even if the actual condition remains unconfirmed.

Walsh [206] and Umemneku Chikere et al. [207] made similar categorisations in their respective literature reviews.

The literature reviews of Umemneku Chikere et al. [207, 208] highlight a number of different approaches that have been suggested to correct for no gold standard. Since the methods presented by Staquet et al. [209] and Gart and Buck [210] are equivalent and shown to produce senstivity and specificity estimates greater than 1, they are not considered here. We will then consider several methods that attempt to calculate sensitivity, specificity and positive predictive value where there is an imperfect gold standard before suggesting a new method that makes use of imprecise probabilities to characterise the uncertainty.

This chapter continues as follows: first the statistics that are used to assess the performance of binary classification tests are reviewed. Section 4.2 explores an approach suggested by Winkler and Smith [211] that presents a method to characterise the performance of a test on a specific individual. Section 4.3 presents a novel method using imprecise probabilities to characterise the uncertainty. Sections 4.5 and 4.6 review methods presented within the literature. Finally there is a short discussion of the problem and the approaches explored within this chapter.

## 4.1 What Makes a 'Good' Test?

Let $T$ be a test for condition $D$. To assess whether $T$ is good at classifying between those with $D$ and those without $D$ ($\neg D$), there are two important statistics that can be used:

- **Sensitivity** ($s$) - Out of those who actually have $D$, the fraction that received a positive test result ($N_{T+}$):

$$s = \frac{\text{Number of True Positive}}{\text{Total Number of Positives}}. \tag{4.1}$$

- **Specificity** ($t$) - Out of those who did not have $D$ ($\neg D$), the fraction that received a negative test result ($N_{T-}$):

$$t = \frac{\text{Number of True Negative}}{\text{Total Number of Negatives}}. \tag{4.2}$$

In order to obtain these values, we need to use the test on a population for which the actual classifications are known. Suppose we have $N$ samples, of which $X$ have $D$, and $Y$ do not. We can use $T$ on these $N$ samples, returning $a$ true positives, $b$ false positives, $c$ false negatives and $d$ true negatives as shown in Figure 4.1. These values can be tabulated into a confusion matrix as shown in Table 4.1.

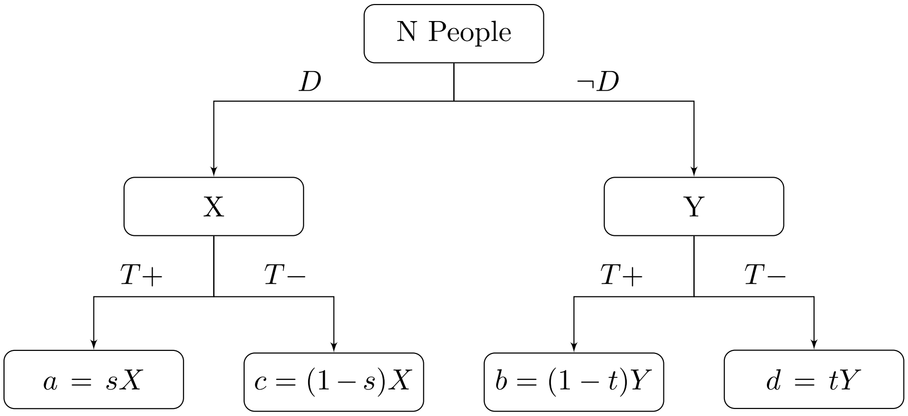


Figure 4.1: Testing for condition $D$ using test $T$.

| | $D$ | $\neg D$ | Total |
|---|---|---|---|
| Tested Positive | $a$ | $b$ | $a + b = N_{T+}$ |
| Tested Negative | $c$ | $d$ | $c + d = N_{T-}$ |
| Total | $a + c = X$ | $b + d = Y$ | $N$ |

Table 4.1: Confusion matrix

From this confusion matrix, the sensitivity can be calculated as

$$s = \frac{a}{X} = \frac{a}{a+c} \tag{4.3}$$

and the specificity as

$$t = \frac{d}{Y} = \frac{d}{b+d}. \tag{4.4}$$

Importantly, these statistics depend only on the test itself and are presumed to not depend on the population the test is to be used upon. As shown in Figure 4.1, if we exclude the impact of any inferential uncertainty, we would expect that $a = sX$, $b = (1-t)Y$, $c = (1-s)Y$ and $d = tY$.

We can also define the positive probability of the test as the ratio of positive tests out of the total number tested:

$$h = \frac{N_{T+}}{N} = \frac{a+b}{a+b+c+d}. \tag{4.5}$$

When the test is used for diagnostic or classification purposes, the population's characteristics become essential for interpreting the test results. To interpret the value of a positive or negative test result, the following statistics must be used:

- **Prevalence** ($p$): the proportion of people in the target population that have $D$:

  $$p = \frac{X}{N} = \frac{a+b}{a+b+c+d} \tag{4.6}$$

  In a Bayesian context, this is the prior probability of having $D$ before being tested.

- **Positive Predictive Value** ($\psi$): the probability that one has $D$ given they have received a positive test result:

  $$\psi = \Pr(D|T+) \tag{4.7}$$

- **Negative Predictive Value** ($\eta$): the probability that one does not have $D$ given they have received a negative test result:

  $$\eta = \Pr(\neg D|T-) \tag{4.8}$$

$\psi$ and $\eta$ depend on the prevalence of $D$ within the population tested. Thus, it is more often than not inappropriate to use Equations 4.7 and 4.8. It is instead better to calculate $\psi$ and $\eta$ using Bayes' rule [212, 213]:

$$\psi = \Pr(D|T+, p, s, t) = \frac{ps}{ps + (1-p)(1-t)}, \tag{4.9}$$

$$\eta = \Pr(\neg D|T-, p, s, t) = \frac{t(1-p)}{t(1-p) + (1-s)p}. \tag{4.10}$$

To illustrate the impact of prevalence on $\psi$, we can look at Figure 4.2. For a test with $s = t = 0.90$, if prevalence $p = 0.05$, then the $\psi \approx 0.32$, implying under a third of positive results are true positives. Figure 4.2a shows that for 1000 test subjects, there will be many more false positives than true positives even with a relatively high sensitivity and specificity of 90%. In contrast, using the same tests on a sample with equal numbers of cases with and without $D$ ($p = 0.5$), we find that $\psi = \eta = 0.9$, see Figure 4.2b. Similarly, the $\eta$ is lower when the prevalence is higher, as shown in Figure 4.2c.

Tests are often repeated to improve the diagnostic performance, or multiple tests are used concurrently to increase the aggregate $\psi$ or $\eta$, assuming that the repeated tests are independent of each other. In such a situation, the $\psi$ of the second test can be calculated using Equation 4.9 but using the $\psi$ of the first test instead of the prevalence.

In a machine learning or information retrieval context, sensitivity is referred to as recall, and the positive predictive value is known as the precision (and calculated using Equation 4.7).

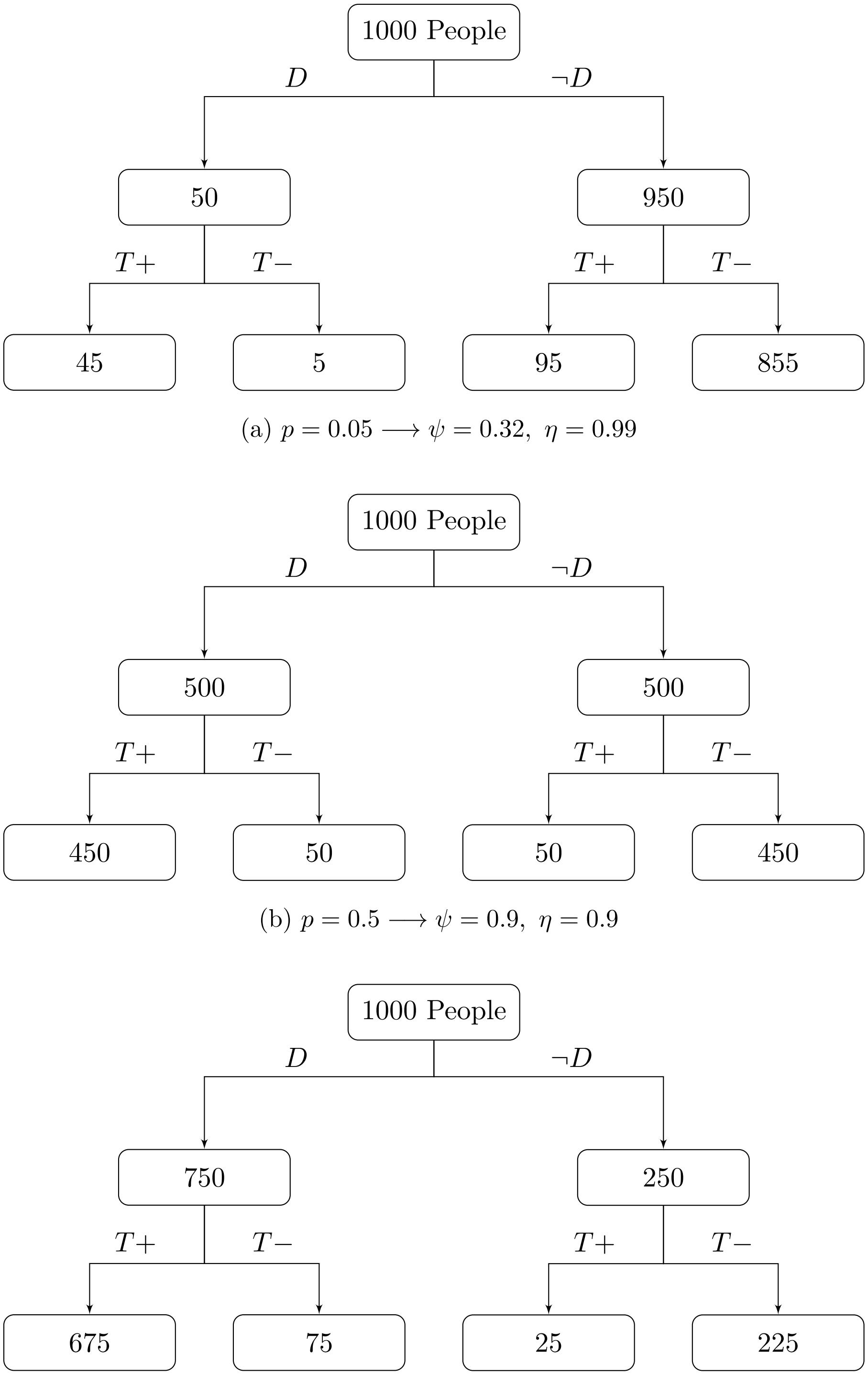
In the medical domain, the values of $p$, $s$ and $t$ are often published independently and come from separate clinical trials. In this case, the following notation can also be used:

$$p = \frac{p_k}{p_n} \tag{4.11a}$$

$$s = \frac{s_k}{s_n} \tag{4.11b}$$

$$t = \frac{t_k}{t_n} \tag{4.11c}$$

where is the number of people in the study and $p_k$ is the number of people who are actually postive. $s_n$ is the number of people with the disease that the test was used on and $s_k$ the number of true results, et cetera for $t_n$ and $t_k$. From the confusion matrix paradigm: $p_k = a + b$, $p_n = a + b + c + d$, $s_k = a$, $s_n = 1 + c$, $t_k = d$ and $t_n = b + d$.



(a) $p = 0.05 \longrightarrow \psi = 0.32,\ \eta = 0.99$

(b) $p = 0.5 \longrightarrow \psi = 0.9,\ \eta = 0.9$

(c) $p = 0.75 \longrightarrow \psi = 0.96,\ \eta = 0.75$

Figure 4.2: Testing different populations with different prevalences for $D$ using test $T$ with $s = t = 0.90$

### 4.1.1 Uncertainty About Test Statisitcs

Whilst it is often the case that these diagnostic statistics are presented as precise values there can be uncertainty associated with $p$, $s$ and $t$. Mossman and Berger [214] consider the following hypothetical situation:

> *Mr Smith has tested positive for disorder D, he asks his doctor the following: "Given the published estimates for prevalence, sensitivity and specificity, what is the 95% confidence interval for my probability of having D given my positive test results and the imprecision in the estimates?"*

When there is uncertainty about the values of $p$, $s$ and $t$, they can be described by distributions [207, 210, 215, 216]. There are a couple of ways $\psi$ can be determined, and the simplest is to estimate it using Equation 4.9 but replacing $p$, $s$ and $t$ with their expected values:

$$\widehat{\psi} = \frac{E(p)E(s)}{E(p)E(s) + (1 - E(p))(1 - E(t))}. \tag{4.12}$$

In order to obtain a distribution for $\psi$, Mossman-Berger use a convolution of the distributions of $p$, $s$ and $t$ within Equation 4.9. In their numerical calculation, they use Monte Carlo and sample random variables from the distributions of $p$, $s$ and $t$ and use Equation 4.9 to find the distribution of $\psi$ making use of Jeffreys prior to,

$$f_j(x|k, n) = \text{beta}(x|k + 1/2, n - k + 1/2), \tag{4.13}$$

to calculate $p$, $s$ and $t$:

$$p(x) = f_j(x|p_k, p_n) \tag{4.14a}$$

$$s(x) = f_j(x|s_k, s_n) \tag{4.14b}$$

$$t(x) = f_j(x|t_k, t_n). \tag{4.14c}$$

They use Jeffreys prior as a non-informative prior distribution for the variables. Other prior distributions could be used [217]. An imprecise probability approach uses $k$-out-of-$n$ c-boxes to characterise the uncertainty (Section 2.5, Equation 2.29). In this instance,

$$p(x) = \text{KN}(x|a + c, a + b + c + d) \text{ or } \text{KN}(x|p_k, p_n) \tag{4.15a}$$

$$s(x) = \text{KN}(x|a, a + c) \text{ or } \text{KN}(x|s_k, s_n) \tag{4.15b}$$

$$t(x) = \text{KN}(x|d, b + d) \text{ or } \text{KN}(x|t_k, t_n). \tag{4.15c}$$

Calculating $\psi$ and $\eta$ using c-boxes is facilitated by the following single-use expressions:

$$\psi = \frac{1}{1 + \left(\frac{1}{s}\right)(1-t)\left(\frac{1}{p} - 1\right)} \tag{4.16}$$

and

$$\eta = \frac{1}{1 + (1-s)\left(\frac{1}{t\left(\frac{1}{p} - 1\right)}\right)}. \tag{4.17}$$

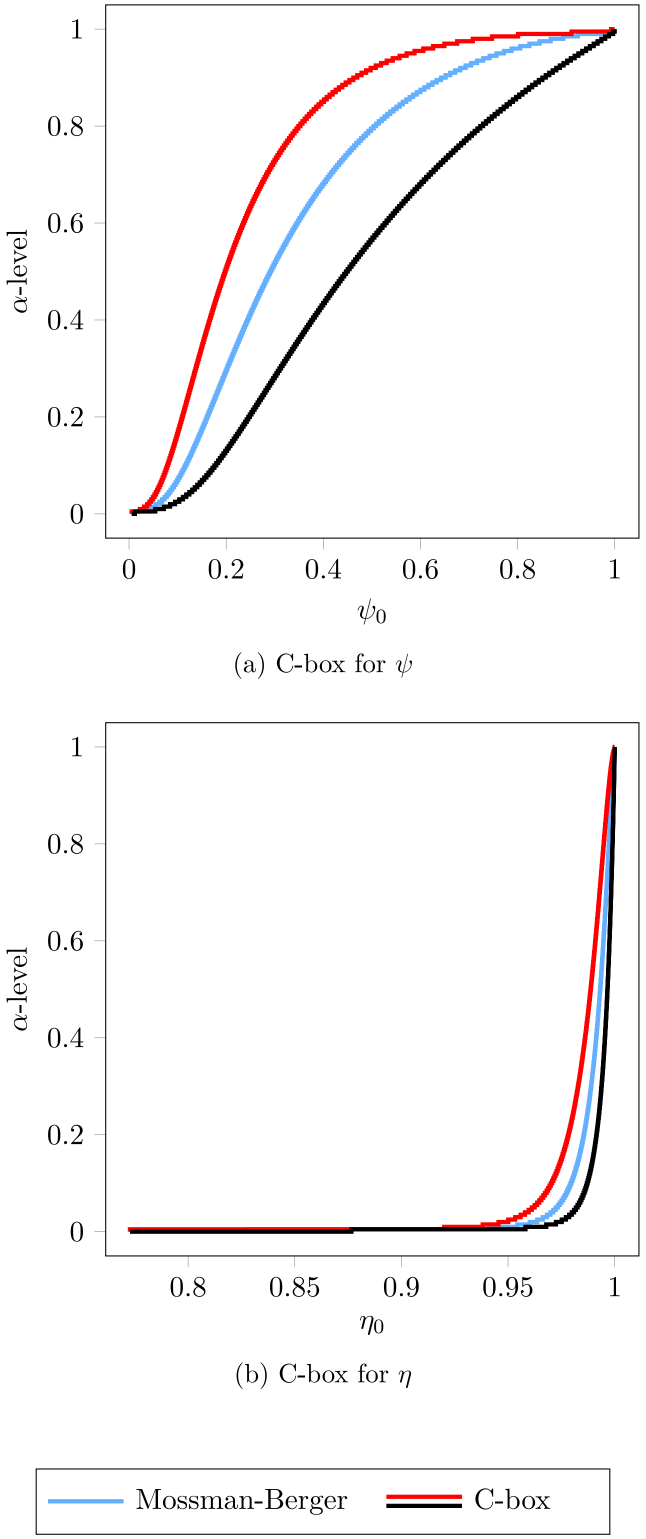
If $p_k = 5$, $p_n = 100$, $s_k = 9$, $s_n = 10$, $t_k = 9$ and $t_n = 10$, where the values of $p$, $s$ and $t$ have come from independent trials, Figure 4.3 shows the distributions the Mossman-Berger (blue) and the c-box method (red/black) estimates for the $\psi$ and $\eta$. As expected the c-box appracoh bounds the Mossman-Berger method. From these plots confidence levels can be drawn. The Mossman-Berger method gives the centered 95% confidence interval for $\psi = [0.07, 0.86]$ and $\eta = [0.97, 1.00]$ and the c-box based approach gives $\psi = [0.05, 0.97]$ and $\eta = [0.96, 1.00]$.

## 4.2 The Winkler-Smith Approach

Winkler and Smith [211] argue that the traditional application of Bayes' rule in medical counselling is inappropriate and represents a "confusion in the medical decision-making literature". They propose in its place a radically different formulation that makes special use of the information about the test results for new patients, although not their actual disease status. Winkler and Smith [211] diverges from Mossman and Berger [214] and the textbook method, see [213] as an example. They argue that the outcome of a patient's test should be used to update the distribution.

As there is no gold standard and the ground truth cannot be immediately established, Winkler and Smith construct two new confusion matrices; one based on the assumption that the test is a true positive and the other assuming the test is a false positive (Table 4.2a and Table 4.2b respectively). They then use these alternatives to update the test's sensitivity, specificity and underlying prevalence. This reducing the test's uncertainty slightly.

Winkler and Smith assert that the positive predictive value of a positive medical test for a disease is not any of Equations 4.9, 4.12 or the Mossman and Berger [214] method (Equation 4.14). Instead, it should be computed as a weighted average of the values calculated using the two new confusion matrices.



(a) C-box for $\psi$

(b) C-box for $\eta$

Figure 4.3: Mossman-Berger confidence distribution and c-box method for calculating $\eta$ and $\psi$ with $p_k = 5$, $p_n = 100$, $s_k = 9$, $s_n = 10$, $t_k = 9$ and $t_n = 10$.

|  | $D$ | $\neg D$ | Total |
|---|---|---|---|
| $T+$ | $a+1$ | $b$ | $N_{T+}$ |
| $T-$ | $c$ | $d$ | $N_{T-}$ |
| Total | $X+1$ | $Y$ | $N$ |

(a) New confusion matrix assuming true positive

|  | $D$ | $\neg D$ | Total |
|---|---|---|---|
| $T+$ | $a$ | $b+1$ | $N_{T+}$ |
| $T-$ | $c$ | $d$ | $N_{T-}$ |
| Total | $X$ | $Y+1$ | $N$ |

(b) New confusion matrix assuming false positive.

Table 4.2: Two new confusion matrices created in the Winkler-Smith method.

Assuming a true positive; the prevalence, sensitivity and specificity become:

$$p'_{+} = \frac{X+1}{N+1} \tag{4.18a}$$

$$s'_{+} = \frac{a+1}{X+1} \tag{4.18b}$$

$$t'_{+} = t. \tag{4.18c}$$

Similarly, assuming a false positive gives:

$$p'_{-} = \frac{X}{N+1} \tag{4.19a}$$

$$s'_{-} = s \tag{4.19b}$$

$$t'_{-} = \frac{d+1}{Y+1}. \tag{4.19c}$$

They then say to construct a weighted average of the new confusion matrices using

$$\widehat{\psi}_{WS} = f(p'_{+}, s'_{+}, t'_{+})\widehat{\psi} + f(p'_{-}, s'_{-}, t'_{-})(1 - \widehat{\psi}), \tag{4.20}$$

where $f(p, s, t)$ is a convolution of the Jeffreys prior distributions (Eq 4.13) for $p$, $s$ and $t$:

$$f(p, s, t) = f_j(p|p_k, p_n)f_j(s|s_k, s_n)f_j(t|t_k, t_n), \tag{4.21}$$

and $\widehat{\psi}$ is calculated using Equation 4.12.

Returning to the numerical example at the end of 4.1.1 ($p_k = 5$, $p_n = 100$, $s_k = 9$, $s_n = 10$, $t_k = 9$ and $t_n = 10$), we find that the 95% confidence interval is $[0.062, 0.741]$. Notice that the width of this confidence interval is smaller than that of the Mossman and Berger method ($\psi_M B = [0.07, 0.86]$ and $\eta_M B = [0.97, 1.00]$)

### 4.2.1 The Logical Inconsistency with the Winkler-Smith Method

To demonstrate that the Winkler and Smith method is flawed, we will explore the situation in which one conducts multiple tests and show that it leads to a *reductio ad absurdum.*

If we consider performing $x$ tests using the Winkler and Smith method, we would have the following two confusion matrices: Table 4.3a and Table 4.3b.

| | $D$ | $\neg D$ | Total |
|---|---|---|---|
| $T+$ | $a+x$ | $b$ | $N_{T+}$ |
| $T-$ | $c$ | $d$ | $N_{T-}$ |
| Total | $X+x$ | $Y$ | $N+x$ |

(a) New confusion matrix assuming true positive

| | $D$ | $\neg D$ | Total |
|---|---|---|---|
| $T+$ | $a$ | $b+x$ | $N_{T+}$ |
| $T-$ | $c$ | $d$ | $N_{T-}$ |
| Total | $X$ | $Y+x$ | $N+x$ |

(b) New confusion matrix assuming false positive.

Table 4.3: Two new confusion matrices created in the Winkler-Smith method with $x$ new results

In the assumed true positive case, the prevalence, sensitivity and specificity become:

$$p'_+ = \frac{X+x}{N+x} \tag{4.22a}$$

$$s'_+ = \frac{a+x}{X+x} \tag{4.22b}$$

$$t'_+ = t, \tag{4.22c}$$

and in the assumed false positive case:

$$p'_- = \frac{X}{N+x} \tag{4.23a}$$

$$s'_- = s \tag{4.23b}$$

$$t'_- = \frac{d+x}{Y+x}. \tag{4.23c}$$

We will now consider what happens to $p'_+$, $s'_+$, $t'_+$, $p'_-$, $s'_-$ and $t'_-$ as $x \to \infty$. We find

$$\lim_{x\to\infty} p'_+ = 1 \tag{4.24a}$$

$$\lim_{x\to\infty} s'_+ = 1 \tag{4.24b}$$

$$\lim_{x\to\infty} t'_+ = t, \tag{4.24c}$$

$$\lim_{x\to\infty} p'_- = 0 \tag{4.25a}$$

$$\lim_{x\to\infty} s'_- = s \tag{4.25b}$$

$$\lim_{x\to\infty} t'_- = 1. \tag{4.25c}$$

Hence, using Equation 4.9 we get

$$\lim_{x\to\infty} \widehat{\psi}_{WS} = 1 \tag{4.26}$$

which implies that $\Pr(\neg D|+) = 0$. If one were to naively interpret this result at face value, they would conclude that any positive test at this limit is a true positive. Winkler and Smith do not say to use this result though their would use Equation 4.20. In this limit, this gives the following:

$$\widehat{\psi}_{WS} = \text{beta}(p|X+x, N+x) \times f_\beta(s|a+x, a+c+x) \times f_\beta(t|d, b+d). \tag{4.27}$$

As

$$\lim_{a,b\to\infty} \text{beta}(x \mid a,b) = \begin{cases} \infty & x = 1/2 \\ 0 & x \neq 1/2 \end{cases} \tag{4.28}$$

and

$$\int_{-\infty}^{\infty} \Pr(x)\mathrm{d}x = 1 \tag{4.29}$$

then the cumulative distribution for $\widehat{\psi}_{WS}$ becomes

$$\widehat{\psi}_{WS} = \begin{cases} 0 & \psi_0 < 1/2 \\ 1 & \psi_0 \geq 1/2 \end{cases}. \tag{4.30}$$

Figure 4.4 shows this migration from the first test, $x = 1$ towards the result in Equation 4.30 starting from the Section 4.1 example data set. These results amount to a logical flaw in the Winkler and Smith method. We have not added any new information (apart from the number of tests), and the uncertainty has been reduced. It should be noted that this asymptote is due to the choice of prior. Interestingly, and perhaps worryingly, different priors would give different values in the limit $x \to \infty$.

### 4.2.2 Imprecise Probability Fix

It is possible to reconsider the argument made by Winkler and Smith using a framework provided by the theory of imprecise probabilities [140, 218–220]. Under this perspective, the prevalence, sensitivity, and specificity can each be updated as prescribed by Winkler and Smith to yield a distribution for the PPV assuming the patient is actually sick (in which case the test was a true positive), and a distribution for the PPV assuming the patient is not sick (in which case the test was a false positive). However, the appropriate synthesis of these two contingent estimates of the PPV is not a weighted mixture as Winkler and Smith conceive it. Instead, because whether the patient is sick or not is what is not known in this problem, an envelope of the two distributions would be more appropriate.

Returning to the numerical example, the envelope of the two contingent distributions yields

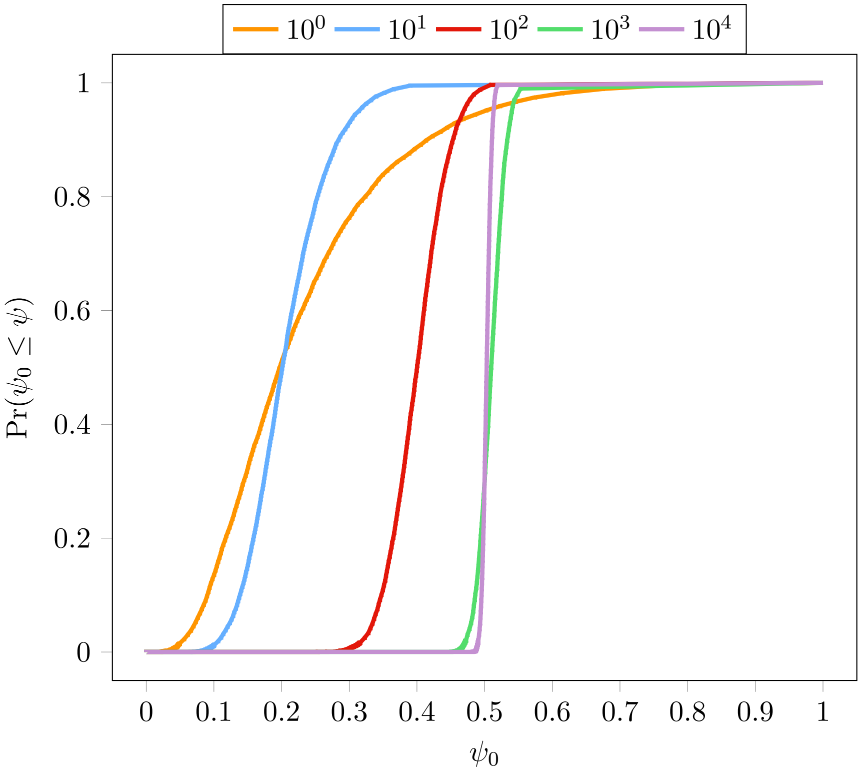


Figure 4.4: Plot of the CDF for the first $10^4$ tests using the Winkler and Smith method.

a rather wide p-box, shown as the outer bounds in blue in Figure 4.5. The leftmost edge corresponds to the distribution that assumes the positive test result was a false positive, incrementing $p$ and $t$ according to Table 4.2b. The rightmost edge corresponds to the distribution that assumes the test result was a true positive, incrementing $p$ and $s$ according to Table 4.2a. This enveloping calculation is equivalent to a mixture with unknown weights characterised by the vacuous interval $[0, 1]$ for both distributions.

In contrast, the traditional Bayesian result and the Winkler-Smith method are also shown in the figure. The envelope encloses both the traditional and the Winkler and Smith distributions. The 95% confidence interval using imprecise probabilities is $[0.057, 0.848]$, which, as expected, encompasses the interval for both the traditional Bayesian result and the Winkler and Smith result. This envelope is reminiscent of the probabilistic dilation of uncertainty that sometimes accompanies the addition of weakly informative data in probabilistic calculations [221].

As we said that the Winkler and Smith methods become logically inconsistent when we consider it in the extreme scenario, we will now show that using imprecise probabilities leads to at least a logical result in the limit. What the imprecise probability confusion

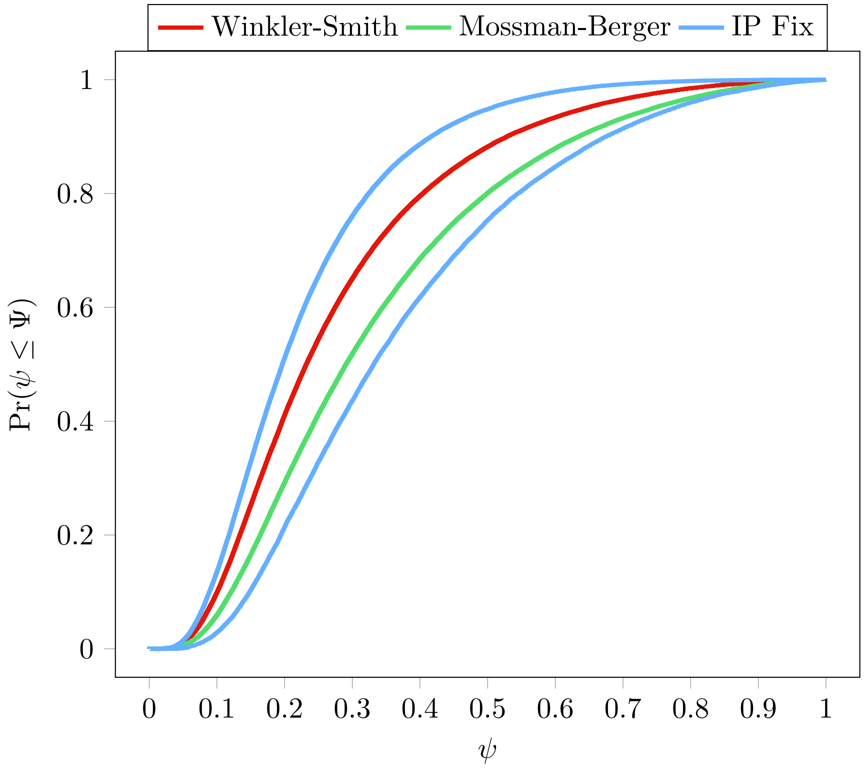


Figure 4.5: P-box showing the distribution envelope for the PPV.

matrix would be after $X$ tests are shown in Table 4.4.

| | Has Problem | No Problem | Total |
|---|---|---|---|
| Positive Test | $a + [0, x]$ | $b + [0, x]$ | $N_{T+} + x$ |
| Negative Test | $c$ | $d$ | $N_{T-}$ |
| Total | $X + [0, x]$ | $Y + [0, x]$ | $N + X$ |

Table 4.4: New imprecise probability confusion matrix after $X$ positive tests.

The new prevalence would be

$$p' = \frac{\alpha + \gamma + X[0, 1]}{T_N + X}, \tag{4.31}$$

new sensitivity

$$s' = \frac{\alpha + X[0, 1]}{\alpha + \gamma + X[0, 1]}, \tag{4.32}$$

and new specificity

$$t' = \frac{\delta}{\beta + \delta + X[0, 1]}. \tag{4.33}$$

Now at the $X \to \infty$ limit:

$$\lim_{X\to\infty} p' = [0,1] \tag{4.34}$$

$$\lim_{X\to\infty} s' = 1 \tag{4.35}$$

$$\lim_{X\to\infty} t' = 0. \tag{4.36}$$

Using these results along with Equation 4.9 gives:

$$\widehat{\psi}_{WS} = [0,1] \tag{4.37}$$

as the final value for the PPV Figure 4.6 shows the migration to the vacuous p-box using the imprecise probability method starting from the numerical example.

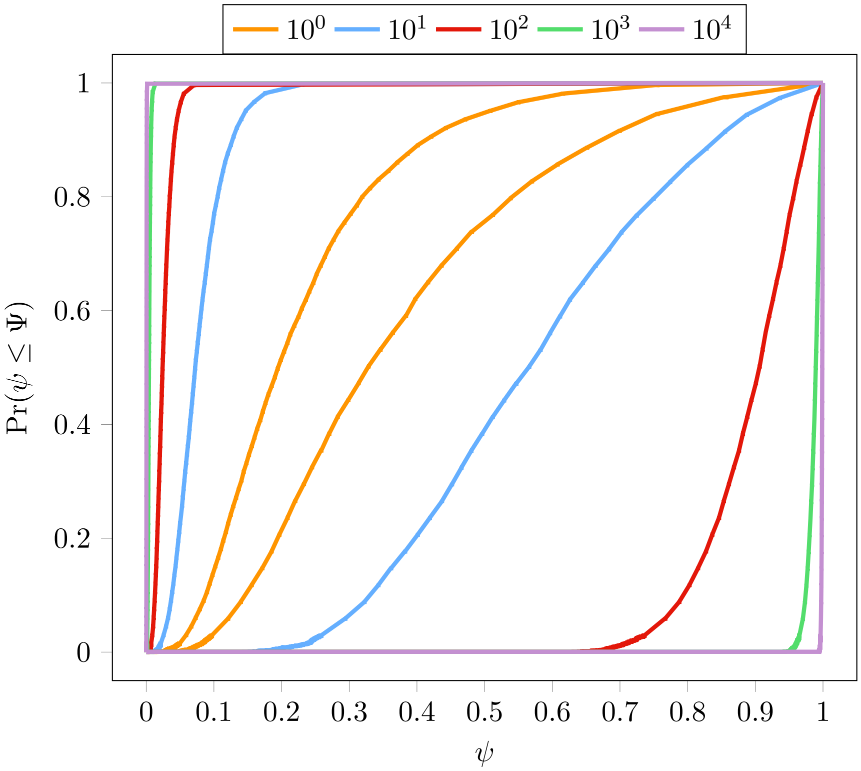


Figure 4.6: Plot of the p-boxes for first 1000 tests using the imprecise probability method.

Let us first consider the difference between the Winkler and Smith method and the imprecise probability method. Figure 4.4 shows that as the number of tests increases, the uncertainty of PPV decreases. This amounts to a *reductio ad absurdum*, thus proving their method untenable. This uncertainty reduction happens even after one test, as demonstrated in the numerical example where the 95% confidence interval from the Winkler-Smith method is narrower than the interval from the Mossman-Berger method.

In the imprecise version, we have also given the test no information, but the uncertainty has increased, which we argue is reasonable. Although at the infinity limit, the vacuous p-box result is not useful when assessing the test's performance, it at least makes logical sense. It is perfectly reasonable to say "I don't know" when one does not know.

## 4.3 A Novel IP Approach

For confusion matrices to be computed, a gold standard test, $G$, that must be able to distinguish between the two classifications perfectly. Without such a test, it would be impossible to know the true classification. In this scenario, only the number of positive and negative tests ($N_{T+}$ and $N_{T-}$) would be known and it would be useful to estimate the number of individuals with $D$ and $\neg D$, $\widehat{X}$ and $\widehat{Y}$ respectively.

Let $\phi_+$ be a measure of how good the positive predictions from the test are, and $\phi_-$ a measure of how good the negative predictions of the test are ($\phi_+, \phi_- \in [0,1]$). Several constraints need to be considered:

**C1** – The estimated values must be equal to the total number tested. i.e.

$$\widehat{X} + \widehat{Y} = N. \tag{4.38}$$

This constraint implies that $\widehat{X}$ and $\widehat{Y}$ are oppositely dependent on each other. As the value of $\widehat{Y}$ increases, the value of $\widehat{B}$ must decrease.

**C2** – As the performance of the test decreases, then the estimates should improve. i.e. If

$$\phi_+ \to 1 \text{ and } \phi_- \to 1 \tag{4.39}$$

then

$$\widehat{X} \to X \tag{4.40}$$

$$\widehat{Y} \to Y. \tag{4.41}$$

**C3** – As the performance of the test decreases, then the estimates should swap. i.e. If

$$\phi_+ \to 0 \text{ and } \phi_- \to 0 \tag{4.42}$$

then

$$\widehat{X} \to Y \tag{4.43}$$

$$\widehat{Y} \to X. \tag{4.44}$$

**C4** – (C2) and (C3) imply that the maximum uncertainty occurs when

$$\phi_+ = \phi_- = \frac{1}{2}. \tag{4.45}$$

At this limit

$$E\left[\widehat{X}\right] = E\left[\widehat{Y}\right] = \frac{N}{2}. \tag{4.46}$$

**C5** – If $\phi_+ \to 1$ but $\phi_- \to 0$ then

$$\widehat{X} \to N \tag{4.47}$$

$$\widehat{Y} \to 0 \tag{4.48}$$

and vice versa.

Using these constraints, we can then construct a function that can *gild* $N_{T+}$ and $N_{T-}$ to produce estimates for $\widehat{X}$ and $\widehat{Y}$ from $\phi_+$ and $\phi_-$.

$$\widehat{X} = g\left(N_{T+}, N_{T-}, \phi_+, \phi_-\right) = N_{T+}\mathrm{KN}(N_{T+}\phi_+, N_{T+}) + N_{T-}\mathrm{KN}(N_{T-}\phi_-, N_{T-}), \tag{4.49}$$

similarly

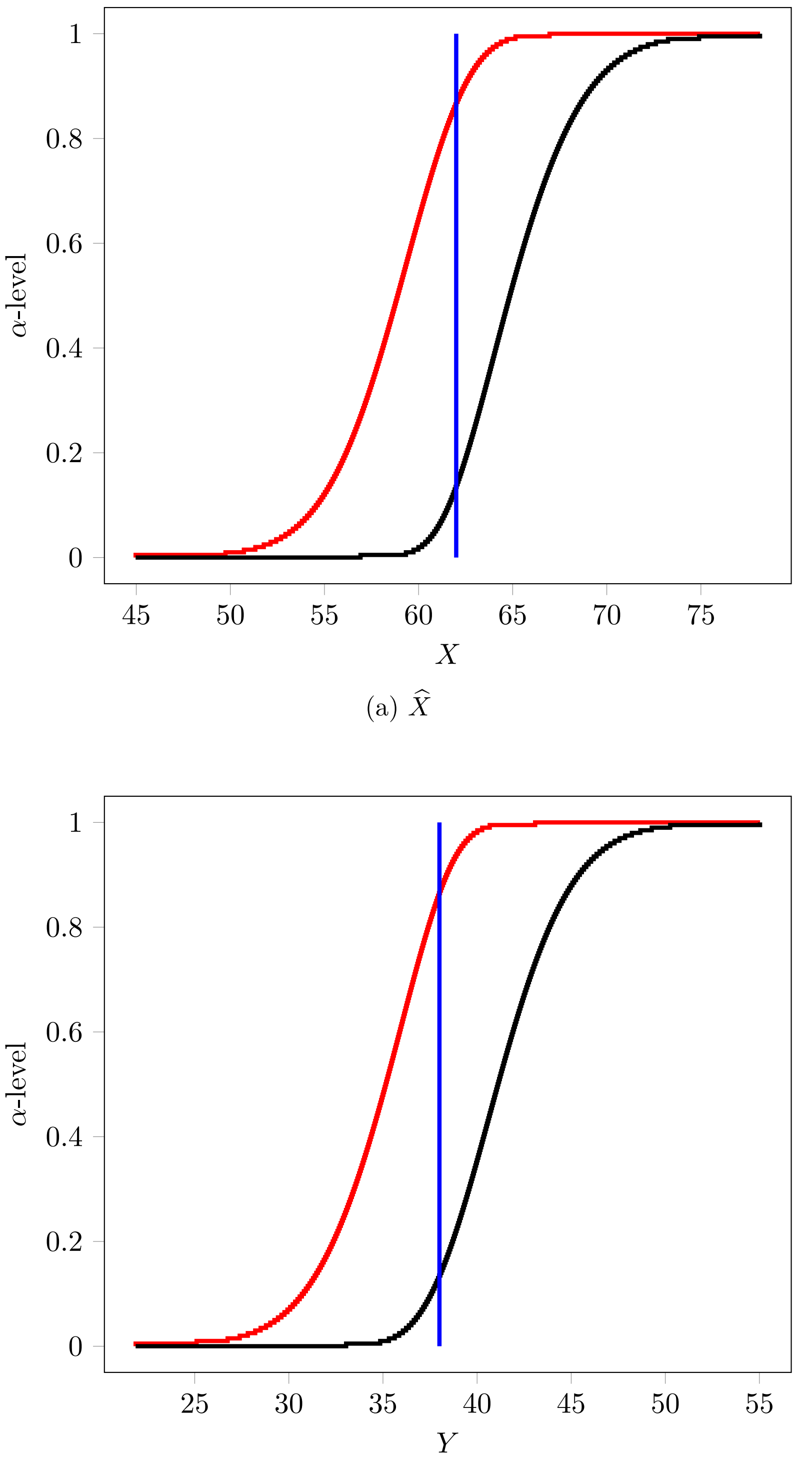
$$\widehat{Y} = g\left(N_{T-}, N_{T+}, \phi_-, \phi_+\right). \tag{4.50}$$

Where KN is the $k$-out-of-$n$ confidence box [158]:

$$\mathrm{KN}(k, n) = \left[\mathrm{beta}(k+1, n-k), \mathrm{beta}(k, n-k+1)\right]. \tag{4.51}$$

### 4.3.1 Example

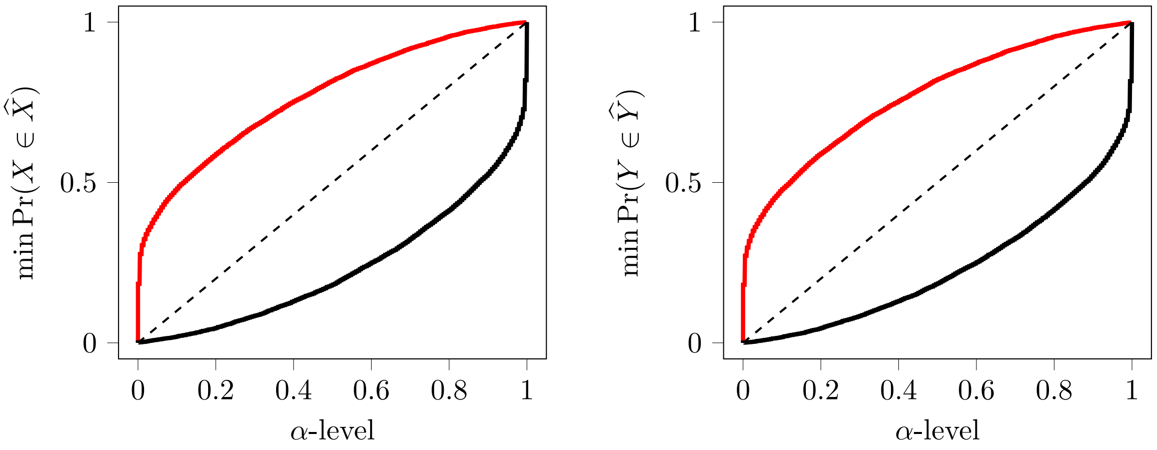
If we have a sample of 100 people that have been tested using test $T$, of which 70 have tested positive and 30, have received a negative result ($N_{T+} = 70$ and $N_{T-} = 30$). The test has $\phi_+ = 0.85$ and $\phi_- = 0.95$. These values can be gilded to produce the c-boxes shown in Figure 4.7. It is useful to compare these c-boxes to the gold standard counts, $X = 62$ and $Y = 38$, which are shown with the blue lines on Figures 4.7a and 4.7b.



(a) $\widehat{X}$

(b) $\widehat{Y}$

Figure 4.7: Gilding c-boxes with $N_{T+} = 70$ and $N_{T-} = 30$ with $\phi_+ = 0.85$ and $\phi_- = 0.95$. The blue lines represents the gold standard count $X = 62$ and $Y = 38$.



(a) $\widehat{X}$.

(b) $\widehat{Y}$.

Figure 4.8: GA-Singh plots for $\widehat{X}$ and $\widehat{Y}$ from Figure 4.7.

In order to assess whether this method performs correctly, it is helpful to consider GA-Singh plots (Section 2.5.1). Figure 4.8 shows the Singh plots for the c-boxes in Figure 4.7. From these plots, we can see that the methodology produces c-boxes that have the confidence interpretation. However, they are conservative, meaning that confidence intervals produced from this methodology are wider than they need for the confidence interpretation to hold.

### 4.3.2 Finding the $\phi$ Values

So far, no attention has been paid to what $\phi_+$ and $\phi_-$ represent, only that they are a measure of the performance of the test. This section discusses various ways they can be found depending on the amount of information known. If the sensitivity and specificity of $T$ are known, as is the prevalence of the population, then $\phi_+$ can be calculated as $\psi$ (Equation 4.9) and $\phi_-$ as $\eta$ (Equation 4.10).

If there is uncertainty about the prevalence, it is possible to provide interval $p$ values (and therefore interval $\phi$ values). In Figure 4.9, the test returns $X = 40$ and $Y = 60$ and has $s = 0.85$ and $t = 0.95$, but there is uncertainty about the prevalence. Table 4.5 shows the prevalence values used within the figure, the resulting $\phi_+$ and $\phi_-$ values, and the 95% confidence intervals. If the sensitivity and specificity are unknown, then it is possible to estimate the $\phi$ values directly.

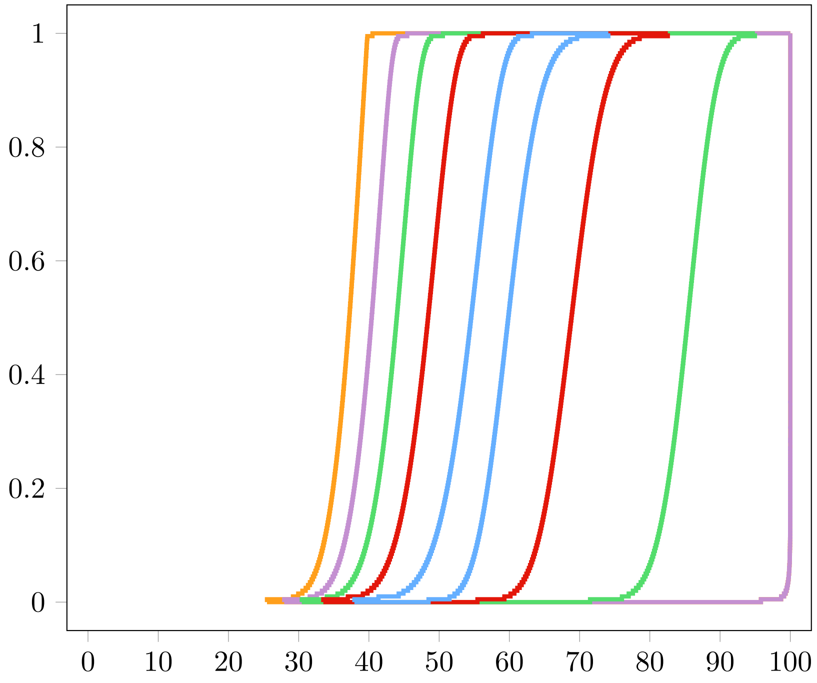


Figure 4.9: Various gilded $X$ values with increasingly wide interval prevalences.

| | $p$ | $\phi_+$ | $\phi_-$ | 95% CI |
|---|---|---|---|---|
| 🟦 | $[0.75, 0.75]$ | $[0.98, 0.98]$ | $[0.68, 0.68]$ | $[46, 68]$ |
| 🟥 | $[0.65, 0.85]$ | $[0.97, 0.99]$ | $[0.53, 0.77]$ | $[41, 77]$ |
| 🟩 | $[0.55, 0.95]$ | $[0.95, 1.00]$ | $[0.25, 0.84]$ | $[37, 91]$ |
| 🟪 | $[0.45, 1.00]$ | $[0.93, 1.00]$ | $[0.00, 0.89]$ | $[34, 100]$ |
| 🟧 | $[0.35, 1.00]$ | $[0.90, 1.00]$ | $[0.00, 0.92]$ | $[32, 100]$ |

Table 4.5: Prvelence, $\phi$ and 95% confidence intervals for $\widehat{X}$ values shown in Figure 4.9.

**Algorithm 4.1:** Algorithm for generating the Singh plots shown in Figure 4.8. $R$ is a random number drawn from the uniform distribution $\mathrm{U}(0,1)$.

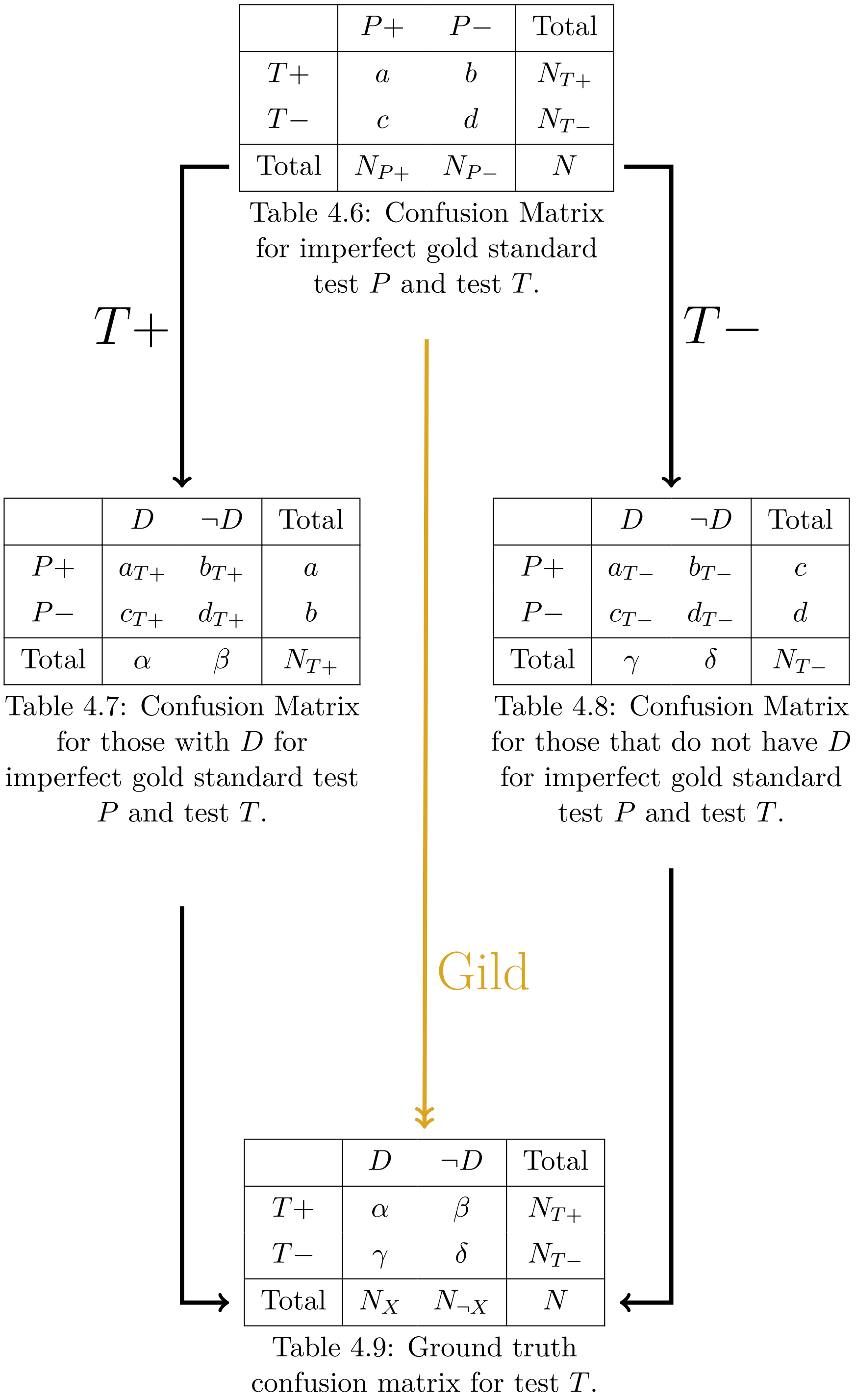
```
Input: $N$, many
for many iterations do
    $p \leftarrow R$
    $\phi_+ \leftarrow R$
    $\phi_- \leftarrow R$
    for all $N$ people do
        TestResult $\leftarrow \{N_{T+}, N_{T-}\}$ $(\Pr(N_{T+}) = p)$
        Status $\leftarrow \{X, \neg X\}$ (If $N_{T+}$ then $\Pr(X) = \phi_+$ else $\Pr(\neg X) = \phi_-$)
    $N_X \leftarrow$ Total $X$
    $Y \leftarrow$ Total $\neg X$
    $N_{N_{T+}} \leftarrow$ Total $N_{T+}$
    $N_{N_{T-}} \leftarrow$ Total $N_{T-}$
    $\widehat{X} = g\left(N_{N_{T+}}, N_{N_{T-}}, \phi_+, \phi_-\right)$
    $\widehat{Y} = g\left(N_{N_{T-}}, N_{N_{T+}}, \phi_-, \phi_+\right)$
    Store $N_X, Y, N_{N_{T+}}, N_{N_{T-}}, \widehat{X}, \widehat{Y}$
Compare all $\widehat{X}$ and $X$ and produce Singh plot;
Compare all $\widehat{Y}$ and $Y$ and produce Singh plot;
```

## 4.4 Pyrite Standards

A potential use of this method is in situations where the performance of a test is measured against an imperfect gold standard (pyrite-standard) test, $P$. In order to find the performance of $T$, we can construct the confusion matrix shown in Table 4.6. Since test $P$ is imperfect, the values of $a$, $b$, $c$ and $d$ and thus $s$ and $t$ calculated from Table 4.6 are incorrect. The true number of true positives, $\alpha$, false positives, $\beta$, false negatives, $\gamma$, and true negatives, $\delta$ are shown in Table 4.9. The true sensitivity and specificity would need to be calculated from this table ($\sigma$ and $\tau$).

In order to find true the performance of $T$, we have to construct the confusion matrix shown in Table 4.9. Since test $P$ is imperfect, this confusion matrix can be split into two separate matrices, one for those individuals that have result $T+$ (Table 4.7) and one for those individuals that have result $T-$ (Table 4.8), as shown in Figure 4.10. The elements of these confusion matrices are unknowable without there being an accessible gold standard.

We can treat Tables 4.7 and 4.8 as two separate instances of the problem described in Section 4.3 and use the method to estimate the values within Table 4.9. These estimates are



| | $P+$ | $P-$ | Total |
|---|---|---|---|
| $T+$ | $a$ | $b$ | $N_{T+}$ |
| $T-$ | $c$ | $d$ | $N_{T-}$ |
| Total | $N_{P+}$ | $N_{P-}$ | $N$ |

Table 4.6: Confusion Matrix for imperfect gold standard test $P$ and test $T$.

| | $D$ | $\neg D$ | Total |
|---|---|---|---|
| $P+$ | $a_{T+}$ | $b_{T+}$ | $a$ |
| $P-$ | $c_{T+}$ | $d_{T+}$ | $b$ |
| Total | $\alpha$ | $\beta$ | $N_{T+}$ |

Table 4.7: Confusion Matrix for those with $D$ for imperfect gold standard test $P$ and test $T$.

| | $D$ | $\neg D$ | Total |
|---|---|---|---|
| $P+$ | $a_{T-}$ | $b_{T-}$ | $c$ |
| $P-$ | $c_{T-}$ | $d_{T-}$ | $d$ |
| Total | $\gamma$ | $\delta$ | $N_{T-}$ |

Table 4.8: Confusion Matrix for those that do not have $D$ for imperfect gold standard test $P$ and test $T$.

| | $D$ | $\neg D$ | Total |
|---|---|---|---|
| $T+$ | $\alpha$ | $\beta$ | $N_{T+}$ |
| $T-$ | $\gamma$ | $\delta$ | $N_{T-}$ |
| Total | $N_X$ | $N_{\neg X}$ | $N$ |

Table 4.9: Ground truth confusion matrix for test $T$.

Figure 4.10: Breaking the $P$–$T$ confusion matrix (Table 4.6) into two confusion matrices based upon the result of test $T$ then recombining into the $D$–$T$ confusion matrix (Table 4.9).

| | $P+$ | $P-$ | Total |
|---|---|---|---|
| $T+$ | 93 | 17 | 110 |
| $T-$ | 12 | 128 | 140 |
| Total | 105 | 145 | 250 |

Table 4.10: Example 1 confusion matrix

$$\widehat{\alpha} = g(a, b, \phi_{++}, \phi_{+-}) \tag{4.52a}$$

$$\widehat{\beta} = g(b, a, \phi_{+-}, \phi_{++}) \tag{4.52b}$$

$$\widehat{\gamma} = g(c, d, \phi_{-+}, \phi_{--}) \tag{4.52c}$$

$$\widehat{\delta} = g(d, c, \phi_{--}, \phi_{-+}) \tag{4.52d}$$

with $g(\dots)$ from Equation 4.49. The sensitivity and specificity of $T$ can then be calculated from the gilded confusion matrix to produce gilded estimates $\widehat{\sigma}$ and $\widehat{\tau}$.

In this case four $\phi$ values need to be found:

- $\phi_{++}$ – A measure of how good the positive predictions of the pyrite test are given that the sample received a positive new test result.
- $\phi_{+-}$ – A measure of how good the negative predictions of the pyrite test are given that the sample received a positive new test result.
- $\phi_{-+}$ – A measure of how good the positive predictions of the pyrite test are given that the sample received a negative new test result.
- $\phi_{--}$ – A measure of how good the negative predictions of the pyrite test are given that the sample received a negative new test result.

$\phi_{++}$ and $\phi_{+-}$ need to be calculated for Table 4.7, similarly $\phi_{-+}$ and $\phi_{--}$ need to be calculated for Table 4.8. For the moment, we shall abstain from commenting on the precise nature of these calculations.

### 4.4.1 Example

In the example 1 confusion matrix shown in Table 4.10, test $T$ has been compared to a pyrite standard test $P$. If we know that $P$ has $\phi_{++} = \phi_{-+} = 0.9$ and $\phi_{+-} = \phi_{--} = 0.75$, then we can gild each element to produce the gilded confusion matrix shown in Table 4.11.

We can use the c-boxes from the gilded confusion matrix to produce c-box estimates for the

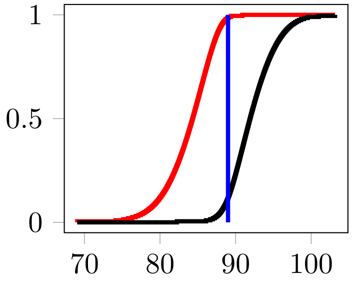
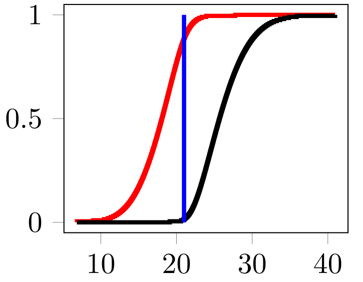
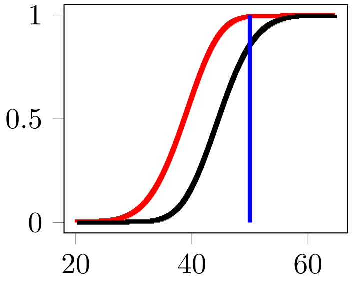
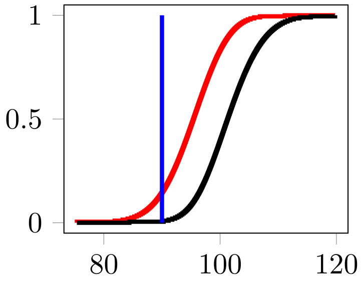
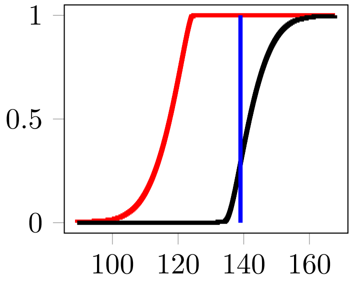
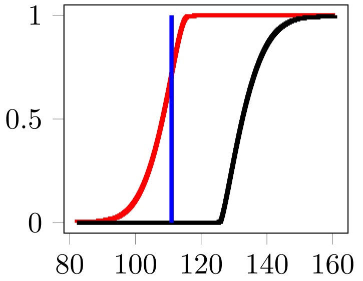
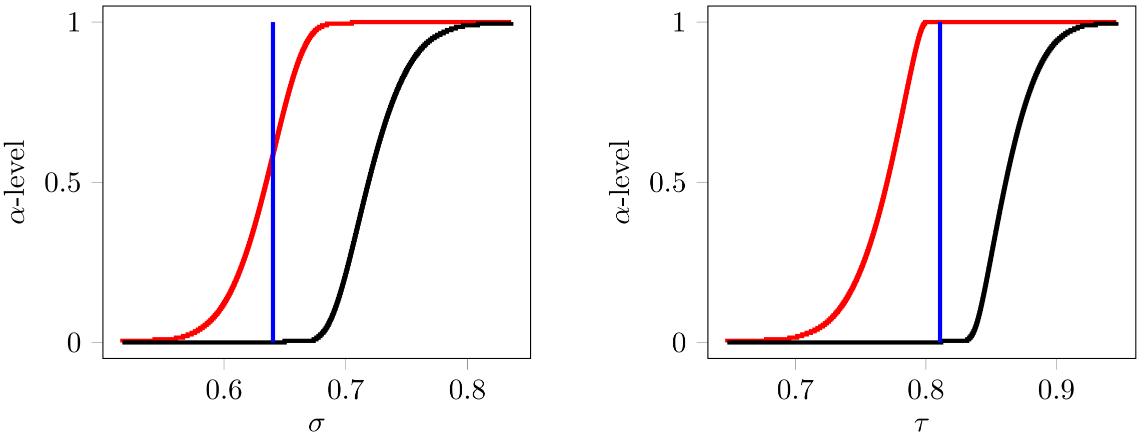
| | $\widehat{X}$ | $\widehat{\neg X}$ | Total |
|---|---|---|---|
| $T+$ |  |  | 110 |
| $T-$ |  |  | 140 |
| Total |  |  | 250 |

Table 4.11: Result of gilding the confusion matrix shown in Table 4.10. The axis labels have been removed from the plots for clarity.



(a) $\widehat{\sigma}$

(b) $\widehat{\tau}$

Figure 4.12: Gilding c-boxes for $\widehat{\sigma}$ and $\widehat{\tau}$ calculated from Table 4.11.

sensitivity, $\widehat{s}$, and specificity, $\widehat{t}$. These are shown in Figure 4.12.

As before, it is helpful to test whether the estimated values have the confidence interpretation using GA-Singh plots. These plots are generated by creating a confusion matrix and randomly sampling $\phi$ values. These plots for $\widehat{\alpha}$, $\widehat{\beta}$, $\widehat{\gamma}$, $\widehat{\delta}$, $\widehat{\sigma}$ and $\widehat{\tau}$ are shown in Figure 4.13. These Singh plots suggest that the presented method, in principle, works when the $\phi$ values are known.

### 4.4.2 The Issue with Prevalence and $\phi$ Values.

In order to produce accurate estimates for $\widehat{\alpha}$, $\widehat{\beta}$, $\widehat{\gamma}$ and $\widehat{\delta}$, it is critical to know the correct $\phi$ values for use within the gilding function. As stated above, when gilding the values within a confusion matrix, 4 $\phi$ values need to be found: $\phi_{++}$, $\phi_{+-}$, $\phi_{-+}$ and $\phi_{--}$.

$\phi_{++}$ and $\phi_{+-}$ are ideally the $\psi$ and $\eta$ of the test for those samples within Table 4.7, similarly for $\phi_{-+}$ and $\phi_{--}$ within Table 4.8. In order to calculate this, we need to find the prevalence of $D$ within Tables 4.7 and 4.8, these are $\Pr(D|T+)$ and $\Pr(D|T-)$ ($\psi$ and $\eta$ of test $T$) respectively. This requires knowledge of $\sigma$ and $\tau$, which we are trying to find.

If assumptions can be made about $\sigma$ and $\tau$, then it may be possible to perform these calculations. For instance, if it is pre-established that the sensitivity and specificity of the new test are at least greater than 0.5 (i.e. the test has some diagnostic value) and the sample prevalence then $\Pr(D|T+)$ and $\Pr(D|T-)$ can be estimated using $\sigma = \tau = [0.5, 1]$, and thus estimated for the prevalence within Tables 4.7 and 4.8. These values can then be used to calculate the appropriate $\phi$ values. For example, starting from the example 2 confusion matrix (Table 4.12), if we know that $p = 0.59$, $s_P = 0.95$ and $t_P = 0.85$ then we

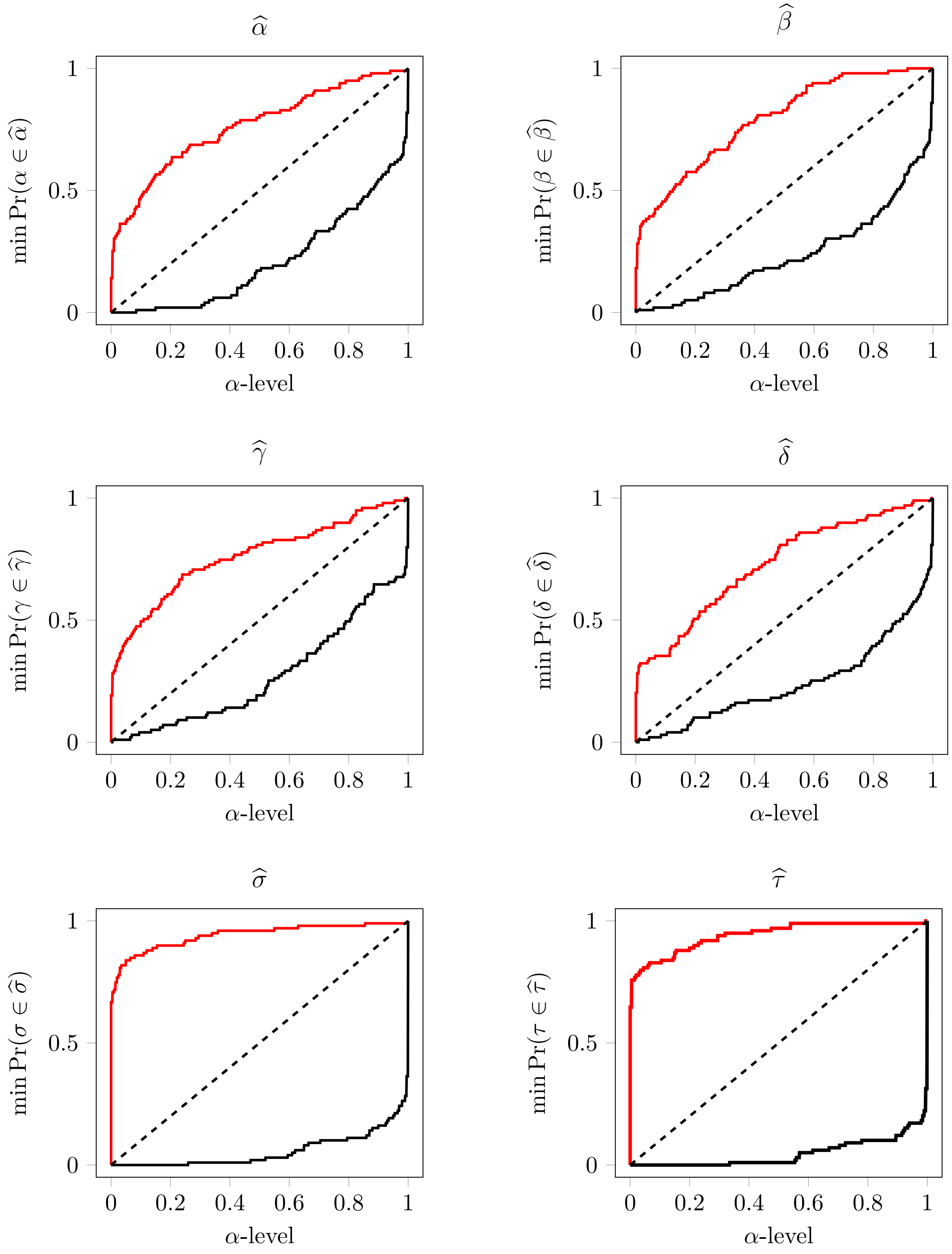
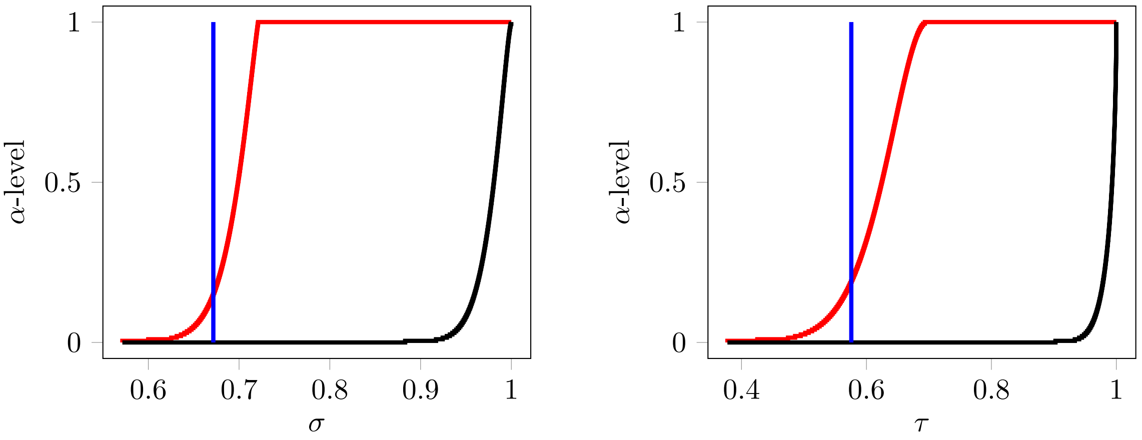


Figure 4.13: GA-Singh plots for $\widehat{\alpha}$, $\widehat{\beta}$, $\widehat{\gamma}$, $\widehat{\delta}$, $\widehat{\sigma}$ and $\widehat{\tau}$.



(a) $\widehat{\sigma}$

(b) $\widehat{\tau}$

Figure 4.15: Gilding c-boxes for $\widehat{\sigma}$ and $\widehat{\tau}$ calculated from Table 4.13.

can create the gilded confusion matrix shown in Table 4.13.

| | $P+$ | $P-$ | Total |
|---|---|---|---|
| $T+$ | 53 | 6 | 59 |
| $T-$ | 16 | 25 | 41 |
| Total | 69 | 31 | 100 |

Table 4.12: Confusion matrix for example 2.

Again we can use GA-Singh plots to test whether the method maintains the confidence interpretation. These plots show that the method shows considerable conservatism, especially about $\widehat{\sigma}$ and $\widehat{\tau}$. If we look at the 95% confidence interval for $\widehat{\sigma}$ and $\widehat{\tau}$ we find $[0.64, 1]$ and $[0.51, 1]$ respectively – almost the values that we started with. This shows that the presented method may be ineffectual unless there is a significant appetite to guess prior values.

## 4.5 The Emerson Approach

Emerson et al. [222] present a method for estimating the sensitivity and specificity of the new test, $T$, from an imperfect gold standard (pyrite standard) test with known sensitivity, and specificity. Their method only works if $T$ and $P$ are not conditionally independent. From the known $t$ and $s$, they calculate the PPV and NPV of the pyrite test using

$$\psi_P = \frac{ps}{h_P} \tag{4.53}$$

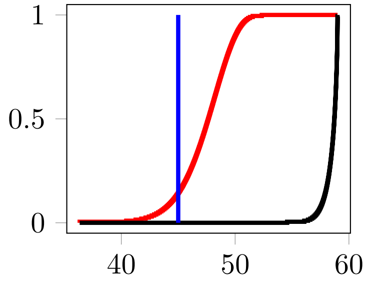
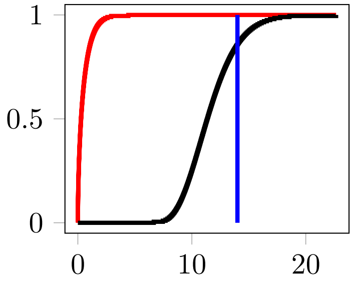
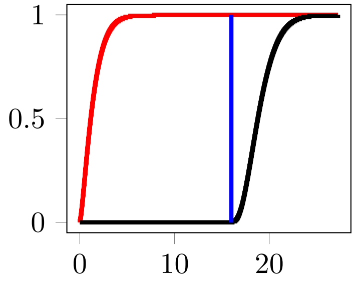
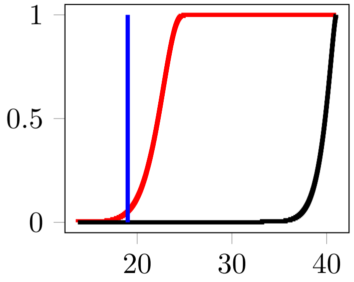
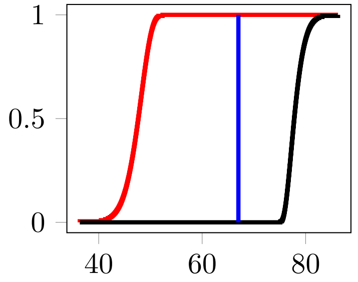
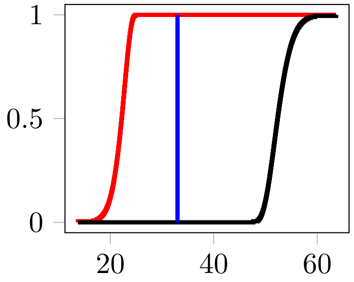
| | $\widehat{X}$ | $\widehat{\neg X}$ | Total |
|---|---|---|---|
| $T+$ |  |  | 59 |
| $T-$ |  |  | 41 |
| Total |  |  | 100 |

Table 4.13: Result of gilding the confusion matrix shown in Table 4.12.

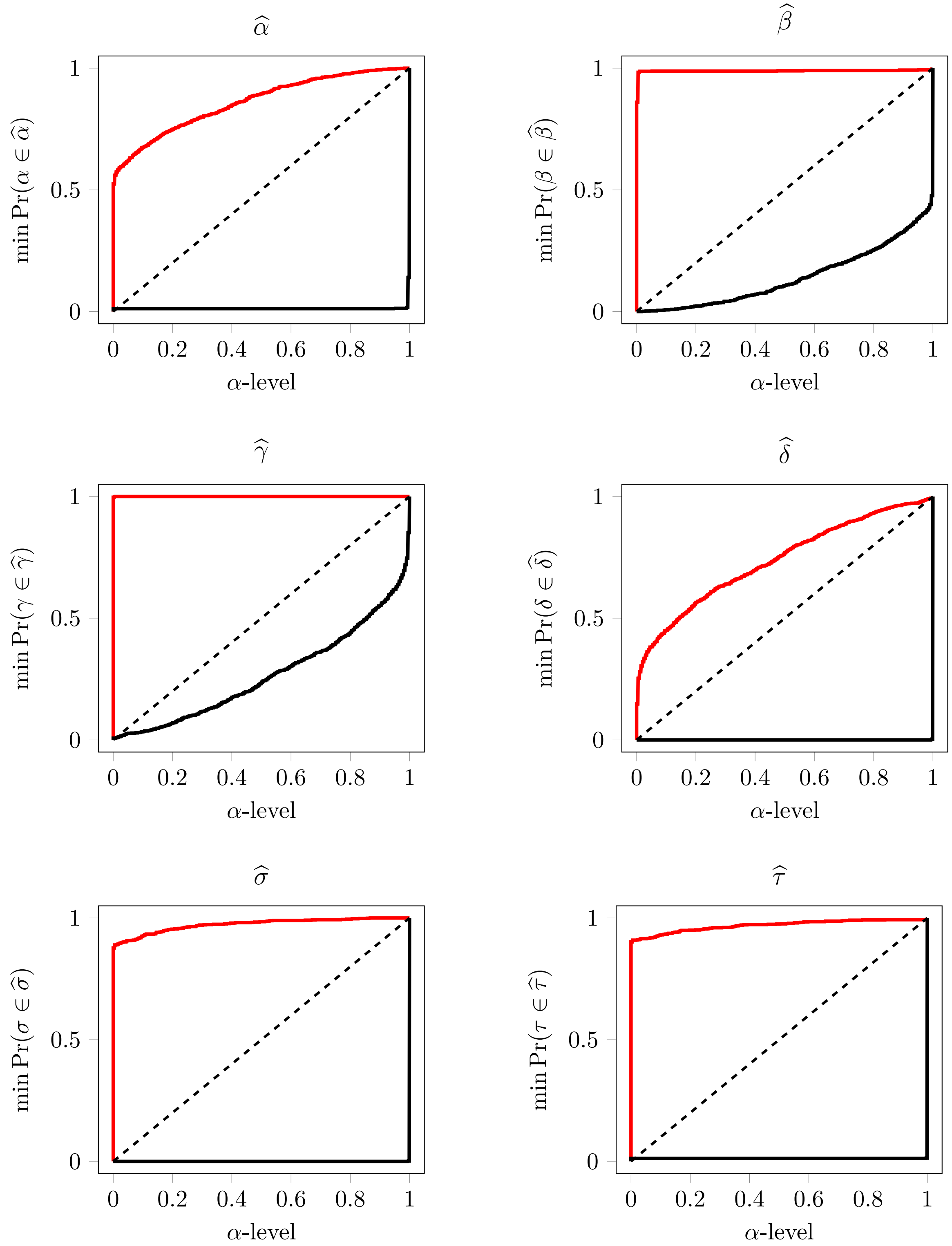


Figure 4.16: GA-Singh plots for $\widehat{\alpha}$, $\widehat{\beta}$, $\widehat{\gamma}$, $\widehat{\delta}$, $\widehat{\sigma}$ and $\widehat{\tau}$.

and

$$\eta_P = \frac{(1-p)t}{1-h_P}, \tag{4.54}$$

where $h_P$ is the probability of testing positive on test $P$, and the prevalence is calculated using

$$p = \frac{h+t-1}{s+t-1}. \tag{4.55}$$

They then calculate the apparent sensitivity, $\sigma'$, and specificity, $\tau'$ of test $T$ (i.e. the values that are calcualted using Table 4.6).

They estimate the true sensitivity and specificity of $T$ as

$$\widehat{\sigma}_E = \frac{\zeta_+ h_P + (1-\eta_P-\tau'+\zeta_-)(1-h_P)}{p} \tag{4.56}$$

and

$$\widehat{\tau}_E = \frac{(1-\psi_P-\sigma'+\zeta_+)h_P + \zeta_-(1-h_P)}{1-p}, \tag{4.57}$$

where $\zeta_{+/-}$ are the positive/negative concordance parameters for which interval bounds can be found

$$\zeta_+ = \Pr(D, T+|P+) = [\max([0, \psi+\sigma'-1]), \min([\psi, \sigma'])] \tag{4.58}$$

$$\zeta_- = \Pr(\neg D, T-|P-) = [\max([0, \eta+\tau'-1]), \min([\eta, \tau'])]. \tag{4.59}$$

Starting with the example 2 confusion matrix (Table 4.12), we can use Equations 4.3, 4.4 and 4.5 to calculate $\sigma' = 0.786$, $\tau' = 0.806$ and $h_P = 0.690$. Equation 4.55 can then be used to calculate $p = 0.675$ and thus Equations 4.53 and 4.54 give $\psi_P = 0.929$ and $\eta_P = 0.891$, hence $\zeta_+ = [0.697, 0.768]$ and $\zeta_- = [0.698, 0.806]$. Finally, giving

$$\widehat{\sigma}_E = [0.713, 0.835]$$

and

$$\widehat{\tau}_E = [0.665, 0.919].$$

Neither of these intervals contain the true values ($\sigma = 0.672$ and $\tau = 0.578$) used to create the sample within the confusion matrix. This is likely because their method does not account for the significant inferential uncertainty associated with the samples.

We can assess the general performance of this method by testing it against a larger number of confusion matrices. If we simulate $10^6$ different confusion matrices for a test with random $\sigma \in [0.5, 0.95]$ and $\tau \in [0.5, 0.85]$, verified against a pyrite standard test with $s = 0.95$ and

$t = 0.85$, with the prevalence being a random number in $[0.05, 0.95]$. We find that

$$\frac{\#(\widehat{\sigma}_E \ni \sigma)}{10^6} = 0.925$$

and

$$\frac{\#(\widehat{\tau}_E \ni \tau)}{10^6} = 0.884$$

.

### 4.5.1 Derivation of $\widehat{\sigma}_E$ and $\widehat{\tau}_E$

Emerson et al. did not provide derivation of their equations for $\widehat{\sigma}_E$ and $\widehat{\tau}_E$. These values are rederived within this section. Instead of breaking the $P$-$T$ confusion matrix (Table 4.6) into two confusion matrices based upon the result of test $T$, Emerson breaks it into confusion matrices based upon the results of test $P$, as shown in Figure 4.17. We can derive $\widehat{\sigma}_E$ and $\widehat{\tau}_E$ from these two tables.

**Derivation of $\widehat{\sigma}_E$** Now

$$\begin{aligned}\sigma &= \Pr(T + |D) \\ &= \frac{\Pr(T + \cap D)}{\Pr(D)},\end{aligned} \tag{4.60}$$

where $\Pr(D)$ is defined as the prevelence ($p$) and

$$\Pr(T + \cap D) = \Pr(D \cap T + \cap P+) + \Pr(D \cap T + \cap P-). \tag{4.61}$$

Since there is conditional dependence between $D$, $P$ and $T$,

$$\Pr(D \cap T + \cap P+) = \Pr(D \cap T + |P+)\Pr(P+) + \Pr(D \cap T + |P-)\Pr(P-). \tag{4.62}$$

From Equations 4.5, $\Pr(P+) = h_P$ and $\Pr(P-) = 1 - h_P$. $\Pr(D \cap T + |P+) = \zeta+$ as defined in Equation 4.58.

In order to find $\Pr(D \cap T + |P-)$ we need to consider that

$$\Pr(D \cap T + |P-) + \Pr(D \cap T - |P-) + \Pr(\neg D \cap T + |P-) + \Pr(\neg D \cap T - |P-) = 1 \tag{4.63}$$

| | $P+$ | $P-$ | Total |
|---|---|---|---|
| $T+$ | $a$ | $b$ | $N_{T+}$ |
| $T-$ | $c$ | $d$ | $N_{T-}$ |
| Total | $N_{P+}$ | $N_{P-}$ | $N$ |

Table 4.6: Confusion matrix for imperfect gold standard test $P$ and test $T$.

$P+$

| | $D$ | $\neg D$ | Total |
|---|---|---|---|
| $T+$ | $\zeta_+\pi_P$ | $(\sigma' - \zeta_+)\pi_P$ | $\sigma'(\pi_P)$ |
| $T-$ | $(\psi_P - \zeta_+)\pi_P$ | $(1 - \psi_P - \sigma' + \zeta_+)\pi_P$ | $1 - \sigma'(\pi_P)$ |
| Total | $\psi_P\pi_P$ | $(1 - \psi_P)\pi_P$ | $N\pi_p$ |

Table 4.14: Confusion matrix for those with $X$ for imperfect gold standard test $P$ and test $T$.

$P-$

| | $D$ | $\neg D$ | Total |
|---|---|---|---|
| $T+$ | $(1 - \eta_P - \tau' + \zeta_-)(1 - \pi_P)$ | $(\eta_p - \zeta_-)(1 - \pi_P)$ | $(1 - \tau')(1 - \pi_P)$ |
| $T-$ | $(\tau' - \zeta_-)(1 - \pi_P)$ | $\zeta_-(1 - \pi_P)$ | $\tau'(1 - \pi_P)$ |
| Total | $(1 - \eta_P)(1 - \pi_P)$ | $\eta_P(1 - \pi_P)$ | $N(1 - \pi_P)$ |

Table 4.15: Confusion matrix for those that do not have $X$ for imperfect gold standard test $P$ and test $T$.

Figure 4.17: Breaking the $P$v$T$ confusion matrix (Table 4.6) into two confusion matrices based upon the result of test $P$.

and by definition $\Pr(\neg D \cap T-|P-) = \zeta_-$. These values are shown within Table 4.14. Now,

$$\begin{aligned} \eta_P &= \Pr(\neg D|P-) \\ &= \Pr(\neg D \cap T+|P-) + \Pr(\neg D \cap T-|P-) \\ &= \Pr(\neg D \cap T+|P-) + \zeta_-, \end{aligned} \tag{4.64}$$

hence

$$\Pr(\neg D \cap T+|P-) = \eta_P - \zeta_-. \tag{4.65}$$

Similarly,

$$\begin{aligned} \tau' &= \Pr(T-|P-) \\ &= \Pr(D \cap T-|P-) + \Pr(\neg D \cap T-|P-) \\ &= \Pr(D \cap T-|P-) + \zeta_-, \end{aligned} \tag{4.66}$$

hence

$$\Pr(D \cap T-|P-) = \tau' - \zeta_-. \tag{4.67}$$

Returning to Equation 4.63 we get

$$\begin{aligned} \Pr(D \cap T+|P-) &= 1 - (\tau' - \zeta_-) - (\eta_P - \zeta_-) - \zeta_- \\ &= (1 - \eta_P - \tau' + \zeta_-). \end{aligned} \tag{4.68}$$

Putting these terms into Equation 4.60 gives:

$$\widehat{\sigma}_E = \frac{\zeta_+ h_P + (1 - \eta_P - \tau' + \zeta_-)(1 - h_P)}{p}. \tag{4.56 ■}$$

**Derivation of $\widehat{\tau}_E$** Now

$$\begin{aligned} \tau &= \Pr(T-|\neg D) \\ &= \frac{\Pr(T-\cap\neg D)}{\Pr(\neg D)}, \end{aligned} \tag{4.69}$$

where $\Pr(\neg D)$ is $(1-p)$ and

$$\Pr(T-\cap\neg D) = \Pr(\neg D \cap T-\cap P+) + \Pr(\neg D \cap T-\cap P-). \tag{4.70}$$

Since there is conditional dependence between $D$, $P$ and $T$,

$$\Pr(\neg D \cap T-\cap P+) = \Pr(\neg D \cap T-|P+)\Pr(P+) + \Pr(\neg D \cap T-|P-)\Pr(P-). \tag{4.71}$$

From Equations 4.5, $\Pr(P+) = h_P$ and $\Pr(P-) = 1 - h_P$. $\Pr(\neg D \cap T - |P-) = \zeta-$ as defined in Equation 4.59.

In order to find $\Pr(\neg D \cap T - |P+)$ we need to consider that

$$\Pr(D \cap T + |P+) + \Pr(D \cap T - |P+) + \Pr(\neg D \cap T + |P+) + \Pr(\neg D \cap T - |P+) = 1, \tag{4.72}$$

and $\Pr(D \cap T + |P+) = \zeta_+$. These values are shown within Table 4.15. Now,

$$\begin{aligned} \psi_P &= \Pr(D|P+) \\ &= \Pr(D \cap T + |P+) + \Pr(\neg D \cap T - |P+) \\ &= \Pr(D \cap T + |P+) + \zeta_- \end{aligned} \tag{4.73}$$

hence

$$\Pr(\neg D \cap T - |P+) = \psi_P - \zeta_+. \tag{4.74}$$

Similarly,

$$\begin{aligned} \sigma' &= \Pr(T + |P+) \\ &= \Pr(D \cap T + |P+) + \Pr(\neg D \cap T + |P+) \\ &= \Pr(D \cap T + |P+) + \zeta_+, \end{aligned} \tag{4.75}$$

hence

$$\Pr(D \cap T + |P+) = \sigma' - \zeta_+. \tag{4.76}$$

Returning to Equation 4.72 we get

$$\begin{aligned} \Pr(\neg D \cap T - |P+) &= 1 - (\sigma' - \zeta_+) - (\psi_P - \zeta_+) - \zeta_+ \\ &= (1 - \psi_P - \sigma' + \zeta_+). \end{aligned} \tag{4.77}$$

Putting these terms into Equation 4.69 gives:

$$\widehat{\tau}_E = \frac{(1 - \psi_P - \sigma' + \zeta_+)h_P + \zeta_-(1 - h_P)}{1 - p}. \tag{4.57 ■}$$

### 4.5.2 Another Imprecise Probability Fix?

In order to account for the inferential uncertainty associated with the sample, it would be useful to use confidence boxes. It would be natural to compute KN c-boxes for $\sigma'$, $\tau'$ and $h_P$. These values could then be used with Emerson's method. Unfortunately, due to a large number of repeated variables, it is impossible to rearrange Equations 4.56 or 4.57

into single-use expressions, as $\psi_P$, $\eta_P$ and $\zeta_{+/-}$ are all dependent on the prevalence. If we assume that the prevalence of the sample is known, then the only c-boxes are $\sigma'$ and $\tau'$.

The artifactual uncertainty can be reduced by reconsidering the values of $\Pr(D \cap T + |P-)$ and $\Pr(\neg D \cap T - |P+)$ used within the derivations.

If we consider the two confusion matrices that Emerson uses within their definitions (Tables 4.14 and 4.15 in Figure 4.17). In each of these tables, the extreme values occur when one of the matrix elements is equal to zero. For instance, in Table 4.15 the extreme value of $\Pr(\neg D \cap T - |P+)$ will occur when either:

- $\Pr(D \cap T + |P+) = \zeta_+ = 0 \implies \Pr(\neg D \cap T - |P+) = 1 - \sigma' - \psi_P$,
- $\Pr(D \cap T - |P+) = \sigma' - \zeta_+ = 0 \implies \Pr(\neg D \cap T - |P+) = 1 - \psi_P$,
- $\Pr(\neg D \cap T + |P+) = \psi - \zeta_+ = 0 \implies \Pr(\neg D \cap T - |P+) = 1 - \sigma'$, or,
- $\Pr(\neg D \cap T - |P+) = 0$.

Hence, because there are imprecise values for $\sigma'$ and $\tau'$,

$$\Pr(\neg D \cap T - |P+) = \operatorname{env}\left(\max(0, 1 - \sigma' - \psi_P), \min(1 - \psi_P, 1 - \sigma')\right) \tag{4.78}$$

and

$$\Pr(D \cap T + |P-) = \operatorname{env}\left(\max(0, 1 - \tau' - \eta_P), \min(1 - \eta_P, 1 - \tau')\right). \tag{4.79}$$

We can use these values to produce new c-boxes for $\widehat{\sigma}_E$ and $\widehat{\tau}_E$, which are shown in Figure 4.18. As can be seen from the plots, this method produces extreme uncertainty about the values. The 95% confidence intervals are $\widehat{\sigma}_E = [0.02, 1.00]$ and $\widehat{\tau}_E = [0.12, 0.99]$. These nearly vacuous intervals imply that there is almost no information that can be gleamed to determine whether the test is good or not. Given the puffiness of the c-boxes, Singh plots have not been created out as such wide confidence intervals would always Singh.

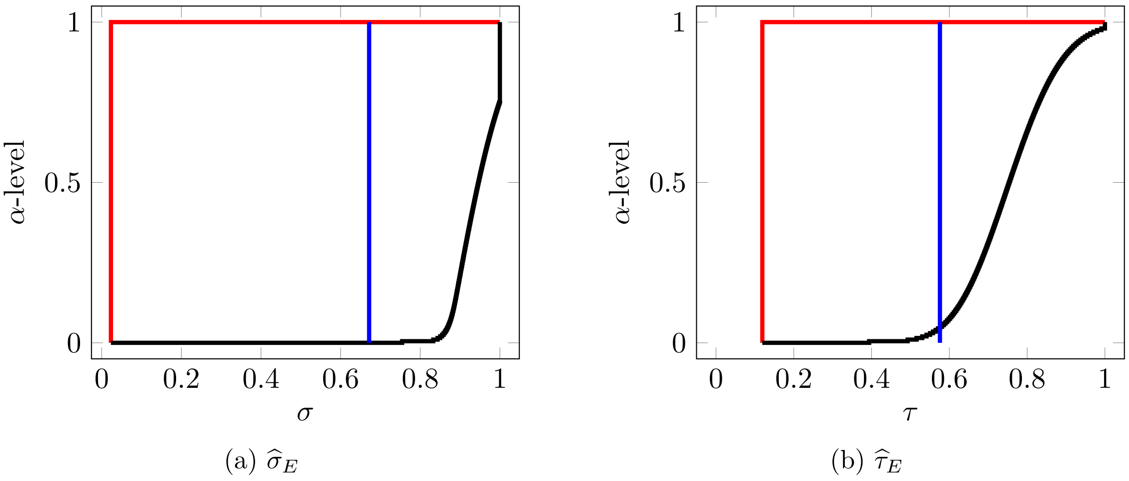
## 4.6 The Brenner Approach

Brenner [223] presents two methods for correcting the sensitivity and specificity with a pyrite standard. Their first method assumes that the two tests are independent, producing:

$$\widehat{\sigma}^i_B = \frac{ps\sigma' + (1-p)(1-t)(1-\tau')}{ps + (1-p)(1-t)} \tag{4.80}$$

and

$$\widehat{\tau}^i_B = \frac{p(1-s)(1-\sigma') + (1-p)t\tau'}{p(1-s) + (1-p)t}. \tag{4.81}$$



(a) $\widehat{\sigma}_E$

(b) $\widehat{\tau}_E$

Figure 4.18: C-boxes produced from the imprecise Emerson method for $\widehat{\sigma}_E$ and $\widehat{\tau}_E$ calculated from Table 4.11.

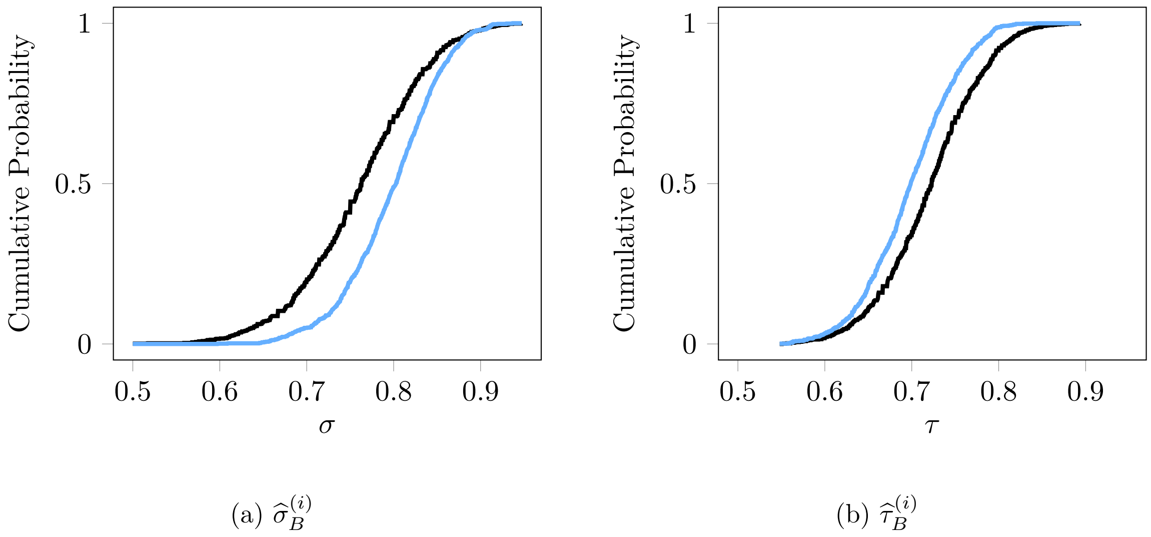
Using these Equations on Table 4.12 (with $p = 0.59$, $s_P = 0.95$ and $t_P = 0.85$) returns $\widehat{\sigma}^i_B = 0.711$, $\widehat{\tau}^i_B = 0.762$. For the comparison $\sigma = 0.672$ and $\tau = 0.578$. Therefore, whilst $\widehat{\sigma}^i_B$ is not dissimilar to the correct value, then $\widehat{\tau}^i_B$ does not appear to be a good estimate.

This singular example does not help assess the method's overall performance. To better understand, it is helpful to simulate 1000 $P$-$T$ confusion matrices, calculate the true values, and use the Brenner method to produce estimates. The empirical cumulative distribution function (ECDF) of these samples can be plotted for both $\sigma$ and $\tau$, as shown in Figure 4.19. From these plots, it is clear that there is a significant difference between the estimates and the true values.

We can then use two-sided two-sample Kolmogorov-Smirnov tests to statistically confirm the lack of similarity between the true values and the estimates. For both of these plots, we find $p \simeq 0^*$ hence it is almost certain that they do not represent the same distribution, and therefore they are poor estimates of the true sensitivity and specificity.

As with the imprecise Emerson case, we can use KN c-boxes to try to improve this method. In this case, replacing $\sigma'$ and $\tau'$ with their c-box equivalents is relatively straightforward. Doing this for the confusion matrix in Table 4.12 gives the c-boxes shown in Figure 4.20. Once again, it is important to test whether these c-boxes maintain the confidence interpretation using GA-Singh plots. Doing this gives the plots shown in Figure 4.21, which clearly shows that the method does not Singh and, therefore, the c-boxes produced do not have the confidence interpretation.

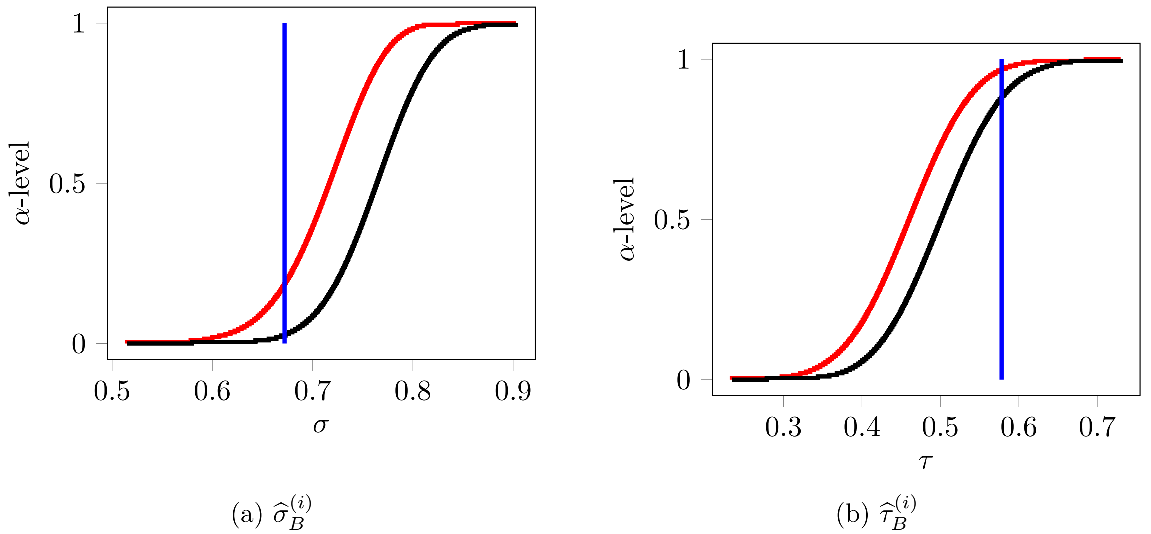
*For $\sigma$ (Figure 4.19a), $p \sim 10^{-15}$ and for $\tau$ (Figure 4.19b) $p \sim 10^{-30}$.



(a) $\widehat{\sigma}_B^{(i)}$

(b) $\widehat{\tau}_B^{(i)}$

Figure 4.19: ECDF for the simulation to test the Brenner approach.



(a) $\widehat{\sigma}_B^{(i)}$

(b) $\widehat{\tau}_B^{(i)}$

Figure 4.20: C-boxes for the imprecise Brenner method.

### 4.6.1 Positive Dependence

Brenner also presents a method for if the two tests are positively correlated. This correlation implies that if the pyrite standard has incorrectly classified an individual, then test $T$ will not correctly classify them. In this scenario, the estimates are:

$$\widehat{\sigma}_B^+ = \frac{p\sigma' + (1-p)(1-t)}{ps + (1-p)(1-t)} \tag{4.82}$$

and

$$\widehat{\tau}_B^+ = \frac{p(1-s) + (1-p)t\tau'}{p(1-s) + (1-p)t}. \tag{4.83}$$

If we consider the confusion matrix in Table 4.16, then using Equations 4.82 and 4.83 we get $\widehat{\sigma}_B^+ = 0.98$ and $\widehat{\tau}_B^+ = 0.88$. These values are significantly different to the gold standard values of $\sigma = 0.743$ and $\tau = 0.721$. Simulating this for 1000 different c-boxes gives the ECDFs shown in Figure 4.22. These two plots differ between the gold standard and the estimates. Therefore the estimates are significantly different to the true values and should be discounted as a way to correct for a pyrite standard test.

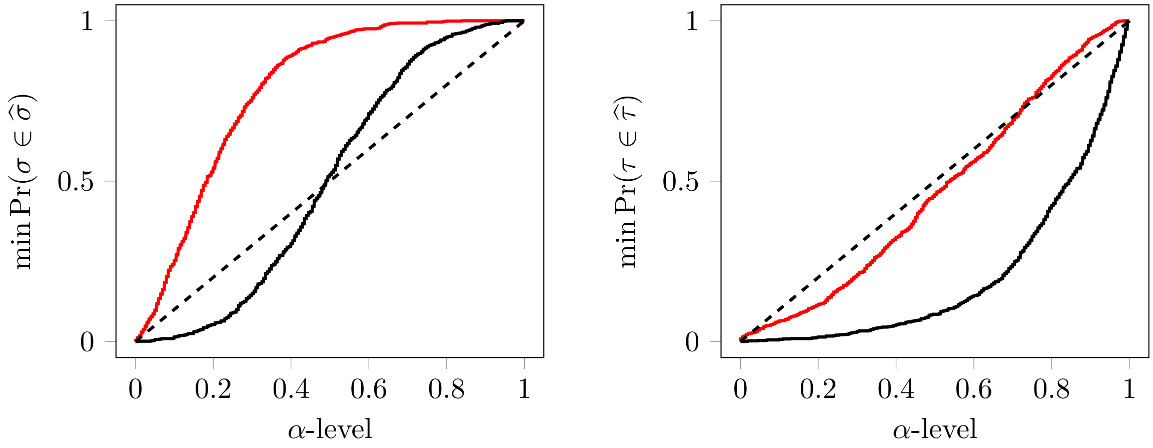
## 4.7 Discussion

The methods described in this chapter show that it is difficult to quantify the epistemic and aleatory uncertainty associated with classification when no gold standard test exists. This is evident as all of the approaches presented struggle to logically, accurately and/or concisely bound the true values. This is a significant problem. Many classification tools are used on data where there is not a gold standard test, as such, the published statistics of these tests may be inaccurate due to errors in the gold standard test that they are compared to.

Winkler-Smith suggested a method to correct the positive predictive value for an individual when there is no gold standard. We have shown that their method is inappropriate, as it leads to the illogical result of the test becoming less uncertain after more trials even though no new information is added. We have shown that it is possible to reimagine their method using imprecise probabilities to create logically consistent results.

The independent and positive correlation approaches suggested by Brenner fail to estimate the actual sensitivity and specificity values and should, therefore, not be used. A c-box based version of their method fails to maintain the confidence interpretation and should equally not be used.

The Emerson method aims to produce interval bounds for the sensitivity and specificity



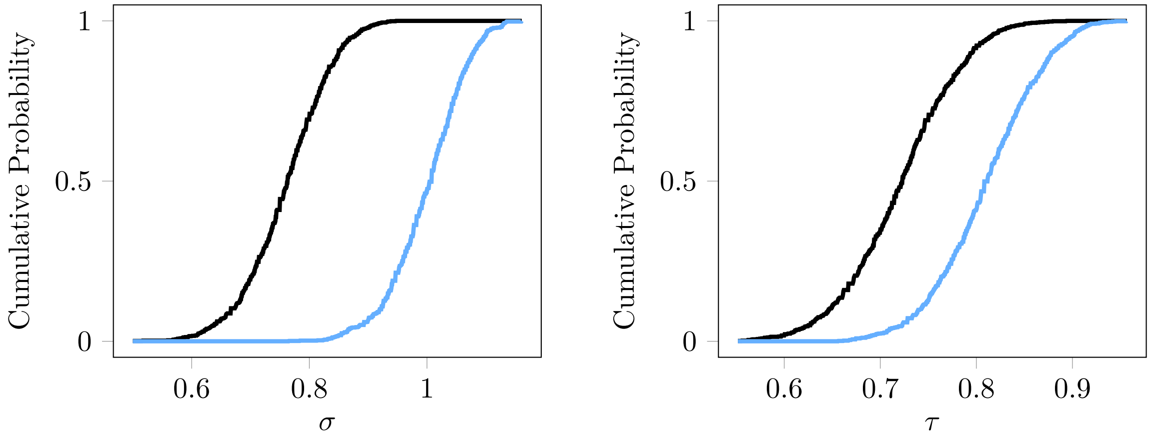
(a) $\widehat{\sigma}_B^{(i)}$

(b) $\widehat{\tau}_B^{(i)}$

Figure 4.21: GA-Singh plots from the imprecise Brenner approach.

| | $P+$ | $P-$ | Total |
|---|---|---|---|
| $T+$ | 36 | 10 | 46 |
| $T-$ | 7 | 47 | 54 |
| Total | 43 | 57 | 100 |

Table 4.16: Confusion matrix for the positively dependent example.



(a) $\widehat{\sigma}_B^{(i)}$

(b) $\widehat{\tau}_B^{(i)}$

Figure 4.22: ECDF for the simulation to test the Brenner approach.

of the test. These bounds were found to contain the true value approximately 90% of the time. Therefore, these intervals have reasonably good coverage of the true value and may be good enough in many scenarios. The imprecise Emerson method produces almost vacuous confidence boxes, unlikely to benefit analysts. Some of this puffiness may result from repeated variables that cannot be reduced within the calculation.

The novel method presented in Section 4.3 produces conservative estimates of the values within the confusion matrix and, therefore, the sensitivity and specificity. The conservatism of the structures shown do not imply that they are necessarily uninformative. For instance, it may be the case that a conservative result can be used to justify that the test is good enough for whatever the use case is.

The issue with the method is that it requires quantifying how good the pyrite test's predictions are at classifying the positive and negative predictions from the new test. The issue with this is that it requires knowledge of the prevalence among those that test positive and negative, which requires producing an estimate of how good the test is.

An alternative approach is to estimate the $\phi$ values via another approach (e.g. by expert elicitation). Or, the method could also be used with various $\phi$ values (or inflating intervals) to see how bad the pyrite test would have to be before the test is deemed unsuitable. One would then use this information to make decisions on the performance of the test.

To conclude, this chapter has shown that finding the true performance of a binary classifier is difficult when there is only a pyrite standard test to verify against.

# Chapter 5

# Interval Uncertainty in Logistic Regression

***Preface***

A pre-print version of this chapter is availible as Nicholas Gray and Scott Ferson. Logistic Regression Through the Veil of Imprecise Data. *arXiv:2106.00492 [stat]*, June 2021. http://arxiv.org/abs/2106.00492.

Logistic regression is used to predict the probability of a binary outcome as a function of some predictive variable. In medicine, for example, logistic regression can be used to predict the probability of an individual having a disease where the values of risk factors are known. While logistic regression is most commonly used for binary outcomes, multinomial logistic regression extends this into events with any number of labels [224]. However, many decisions and events are binary (yes/no, passed/failed, alive/dead, sick/healthy etc.), and for the sake of simplicity, we will restrict our discussion and examples to binary outcome logistic regression. Additionally, unlike discriminant function analysis, logistic regression does not require predictor variables to be normally distributed, linearly related or have equal variance [225].

There are many practical applications for logistic regression across various different fields. For example, in the medical domain, risk factors in the form of continuous data – such as age – or categorical data – such as gender – may be used to fit a model to predict the probability of a patient surviving an operation [226, 227]. In engineering systems, logistic regression can be used to determine whether a mineshaft is safe [228]; and to predict the risk of lightning strikes [229], or landslides [230]. Sports results can be predicted using logistic regression [231]. It was also the method used within the COMPAS recidivism test described in Section 1.3. Due to its wide range of applications, logistic regression is considered a fundamental machine learning algorithm with many modern programming languages having packages for users to dabble with, such as scikit-learn in Python [232], which has been used for the analysis within this chapter.

Traditionally it has been assumed that all of the values of the features and labels used in logistic regression are *precisely* known. However, in practice, there can be considerable imprecision in the features and labels used in the regression analysis and the application of the regression model. Analyses using data from combined studies with inconsistent measurement methods can even result in datasets with varying degrees of uncertainty. Likewise, the outcome data can be uncertain if there is ambiguity in the classification scheme (good/bad). Moreover, even relatively straightforward classifications (alive/dead) can yield uncertainty when a subject leaves a study, and the outcome is unknown. Within statistics, censored data is sometimes of this form [233, 234]. Imprecision might have also been introduced as a way of preserving privacy as discussed in Section 1.6. In the case of continuous variables, the interval reflects the measurement uncertainty, while in the binary outcome, the interval is the vacuous $[0, 1]$ because the correct classification is unclear.

There are multiple methods have been suggested for dealing with interval data with the features of a logistic regression model. This problem may also be considered a subproblem within symbolic data analysis [235–237]. These methods often require approximations to be made to simplify the process and allow the use of standard logistic regression techniques. The most straightforward approach is to neglect the interval data, assuming that the epistemic uncertainty that the intervals represent is small compared to the sampling uncertainty or natural variability in the data, or that values are missing at random or missing completely at random [62, 152, 238, 239]. These assumptions are likely to be untenable in practice. Another approach is to treat interval data as uniform distributions based on the "equidistribution hypothesis" that assumes each possible value to be equally likely; thus, the intervals are modelled as a uniform distribution [237, 240–242]. This idea has its roots in the principle of insufficient reason, first described by both Bernoulli and Laplace, and more recently known as the principle of indifference [243]. Alternatively, the interval is commonly represented by the interval's midpoint, which represents the mean and median of a uniform distribution or a random value from within the interval [61]. While these approaches are computationally expedient, Section 5.2.3 will show, they underrepresent the imprecision by presenting a single middle-of-the-road logistic regression.

Other methods include performing a conjoint logistic regression using the interval endpoints or averaging separate regressions performed on the endpoints of the intervals [244, 245]. While these various methods make different assumptions about the data within the interval ranges, ultimately, they still transform interval data such that the final results can be represented by a single binary logistic regression [245].

The approach proposed within this chapter for dealing with interval data in logistic regression is based on imprecise probabilities and considers a set possible of models rather than a single one [218, 239, 246, 247]. This is similar to approaches proposed for dealing with interval uncertainty within linear regression [248–251]. These approaches, alongside other methods for interval linear regression [136, 251, 252] are not directly translatable to logistic regression as they require the use of least squares methods ill-suited to dichotomous problems [62].

If separate logistic regressions are generated via maximum likelihood estimation from the interval data and displayed graphically, the envelope of these models can be considered as an imprecise logistic regression model. The ‘true’ model – the model that would have been fitted if there were no epistemic uncertainty associated with the sample – would always be contained within the bounds of the best possible imprecise model. The primary benefit of such an approach is that it represents the epistemic uncertainty removed by traditional methods. Additionally, this method can also handle the case of intervals in discrete risk factors. This imprecise approach makes the fewest assumptions but can be computationally challenging for large datasets [152, 239, 253].

In the case of uncertainty in the outcome status used within logistic regression, traditionally there is little that can be done but to discard these data points as they cannot be used as part of the analysis or to use a semi-supervised learning methodology [254–256]. This approach can lead to strongly biased results when the data is not missing at random. However, the proposed imprecise logistic regression technique can be used to include unlabelled examples within the dataset. Again the imprecise approach does not require making the assumptions required by other methods to fit a model.

This chapter continues as follows: in Section 5.1, tradiaitonal precise logistic regression is reviewed. Sections 5.2 and 5.3 introduce the imprecise logistic regression for data with intervals within the features and labels respectively; in both these sections, a 1-dimensional synthetic dataset is used to demonstrate the methodology before it is compared to alternative methods on both the synthetic data and a real-world example. Finally, in Section 5.4 the method is used on a dataset which contains intervals in both the features and labels and is contrasted against a dataset from the literature.

## 5.1 Precise Logistic Regression

Let $\mathbf{x}$ be a $m$-dimensional covariate with a binary label, $y \in \{1, 0\}$. Logistic regression can be used to model the probability that $y = 1$ using:

$$\Pr(y = 1|\mathbf{x}) = \pi(\mathbf{x}) = \frac{1}{1 + \exp\left(-\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m\right)\right)} \tag{5.1}$$

where $\beta_0, \beta_1, \ldots$ are a set of unknown regression coefficients. Dataset $D$ contains $n$ samples:

$$D = \left\{ \begin{array}{c} \left((x_1^{(1)} x_2^{(1)} \cdots x_m^{(1)}), y_1\right) \\ \left((x_1^{(2)} x_2^{(2)} \cdots x_m^{(2)}), y_2\right) \\ \vdots \\ \left((x_1^{(n)} x_2^{(n)} \cdots x_m^{(n)}), y_n\right) \end{array} \right\}. \tag{5.2}$$

Let $\mathcal{LR}(D)$ be a logistic regression model fitted on $D$. This fitting is acheived be finding the optimal values of $\beta_0, \beta_1, \ldots, \beta_m$ such that $\mathcal{LR}(D)$ best fits the data in $D$. This is often done using *maximum likelihood estimation* [224, 257], although other techniques exist, for instance through Bayesian analysis [258, 259].

A classification, $\hat{y}$, can be made from the logistic regression model by selecting a threshold value, $C$, and then defining

$$\hat{y} = \begin{cases} 1 & \text{if } \pi(\mathbf{x}) \geq C \\ 0 & \text{if } \pi(\mathbf{x}) < C \end{cases}. \tag{5.3}$$

The simplest case is when $C = 0.5$, implying $\hat{y}$ is more likely to be true than false. However, this value could be different depending on the use of the model and the risk appetite of the analyst. For example, in medicine, a small threshold value may be used in order to produce a conservative classification and therefore reduce the number of false negative results. Where predictions are made within this chapter, $C = 0.5$ unless otherwise stated.

### 5.1.1 Demonstration

To demonstrate, a synthetic 1-dimensional dataset ($D_1$) with a sample size of fifty was used to train a logistic regression model, $\mathcal{LR}(D_1)$, as shown in Figure 5.1. After training the model, it is useful to ask the question, "how good is the model?" For logistic regression there are several ways in which that can be done, see Hosmer Jr et al. [62, pp. 157–169] or Kleinbaum and Klein [260, pp.318–326]. For the analysis in this paper, we will consider

| | 1 | 0 | Total |
|---|---|---|---|
| Predicted 1 | 34 | 14 | 48 |
| Predicted 0 | 10 | 42 | 52 |
| Total | 44 | 56 | 100 |

Table 5.1: Confusion matrix for 100 test data points using predictions from $\mathcal{LR}(X)$ shown in Figure 5.1.

the receiver operating characteristic graph and area under curve statistic, discriminatory performance visualisations [261], as well as the sensitivity and specificity of the classifications made by the algorithm.

Royston and Altman [261] introduced visualisations to assess the discriminatory performance of the model by considering a scatter plot of the true outcome (jittered for clarity) versus the estimated probability. Such a plot is shown in Figure 5.2a. A perfectly discriminating model would have two singularities with all the points with outcome 1 at (1,1) and all the points with outcome 0 at (0,0). In general, the better the classifier, the more clustered the points would be towards these values, with the points on the upper band having larger probabilities and the points on the lower band having smaller probabilities. From Figure 5.2a we can see that there is significant clustering towards the endpoints, showing that the model has excellent discriminatory performance.

We can make and compare the predictions made from the logistic regression model using a larger dataset that has been generated using the same method as above. Tabulating these results in a confusion matrix for the base predictions is shown in Table 5.1. There are numerous statistics than can be derived from confusion matrices to express the performance of classifiers. For this analysis, we shall consider the sensitivity, $s$, and specificity, $t$ (See Chapter 4, Section 4.1).

From Table 5.1 we can calculate that $s = 0.773$ and $t = 0.750$. As confusion matrices and statistics are calculated from them depending on the cutoff value chosen ($C$ from Equation 5.3), a complete way of determining the classification performance of models is by considering the receiver operating characteristic (ROC) curve of the model [62, 260]. The ROC can be plotted by calculating how the sensitivity and specificity change for various threshold values and then plotting a graph of the false positive rate, $fpr = 1 - t$, against $s$ for all $C$ values. For the example, the ROC curve for $\mathcal{LR}(D)$ is shown in Figure 5.2b. The more upper-left a curve is, the better the classification. The worst performing model's ROC curve would match the black dotted line ($s = fpr$), corresponding to the ROC curve for a random classifier. If a model had a ROC curve down-right of this line, this implies that the performance would be improved by switching the outcome classes of the model,

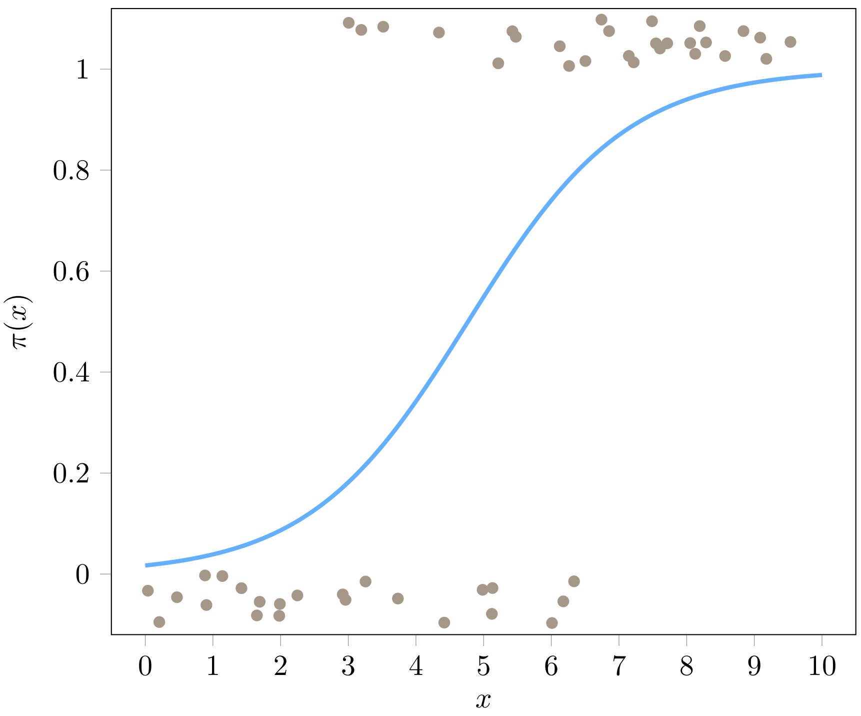
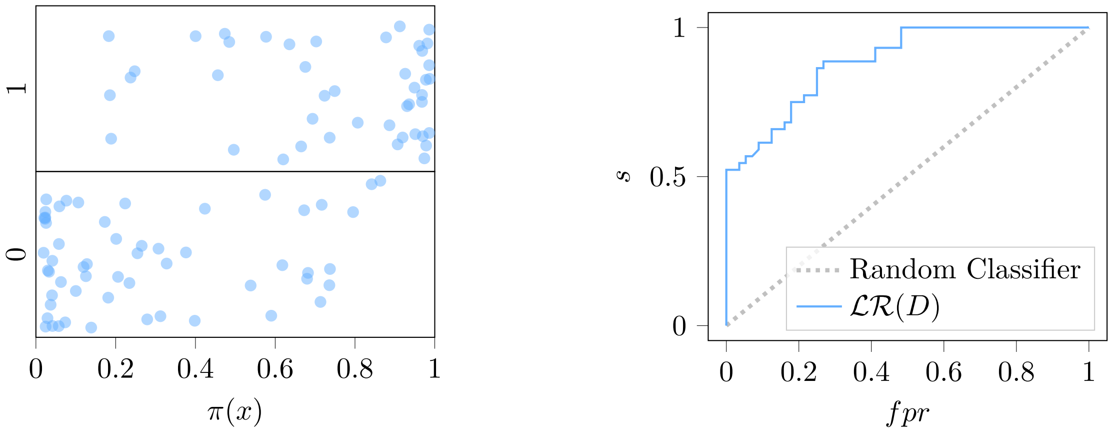


Figure 5.1: Logistic regression curve, $\mathcal{LR}\,(D_1)$, created by fitting the model on dataset $D_1$ (points jittered for clarity).



(a) Scatter plot of jittered outcome vs estimated probability.

(b) ROC curve.

Figure 5.2: Two plots showing the discriminatory performance of the logistic regression model shown in Figure 5.1.

as if it predicted true, then it is more likely to be false and vice versa. ROC curves can be compared graphically and by considering the area under the curve (AUC). The better the model is, the closer the AUC would be to 1. The worst possible AUC would be 0.5, as again, anything lower than that would be improved by simply switching the classification. For the ROC curve shown in Figure 5.2b, AUC = 0.887.

## 5.2 Uncertainty in Features

If there is interval uncertainty within dataset $D$,

$$D = \left\{ \begin{array}{c} \left(\left(\left[\underline{x_1^{(1)}}, \overline{x_1^{(1)}}\right] \left[\underline{x_2^{(1)}}, \overline{x_2^{(1)}}\right] \cdots \left[\underline{x_m^{(1)}}, \overline{x_m^{(1)}}\right]\right), y_1\right) \\ \left(\left(\left[\underline{x_1^{(2)}}, \overline{x_1^{(2)}}\right] \left[\underline{x_2^{(2)}}, \overline{x_2^{(2)}}\right] \cdots \left[\underline{x_m^{(2)}}, \overline{x_m^{(2)}}\right]\right), y_2\right) \\ \vdots \\ \left(\left(\left[\underline{x_1^{(n)}}, \overline{x_1^{(n)}}\right] \left[\underline{x_2^{(n)}}, \overline{x_2^{(n)}}\right] \cdots \left[\underline{x_m^{(n)}}, \overline{x_m^{(n)}}\right]\right), y_n\right) \end{array} \right\}, \tag{5.4}$$

and we have no more information about the true value, $x_j^{(i)\dagger}$, nor we are willing to make further assumptions about the true value, only that $x_j^{(i)\dagger} \in \left[\underline{x_j^{(i)}}, \overline{x_j^{(i)}}\right]$. Then, it is only possible to partially identify an imprecise logistic regression model for the data, $\mathcal{ILR}(D)$:

$$\mathcal{ILR}(D) = \left\{ \mathcal{LR}\left(D'\right) : \forall D' \in \left\{ \left\{ \begin{array}{c} \left(\left(x_1'^{(1)} \cdots x_1'^{(m)}\right), y_1\right) \\ \vdots \\ \left(\left(x_n'^{(1)} \cdots x_n'^{(m)}\right), y_n\right) \end{array} \right\} \forall x_j'^{(i)} \in \left[\underline{x_j^{(i)}}, \overline{x_j^{(i)}}\right] \right\} \right\} \tag{5.5}$$

i.e. $\mathcal{ILR}(D)$ is the set of all possible logistic regression models that can be created from all possible datasets that can be constructed from the interval data. This ensures that the true logistic regression model, $\mathcal{LR}^{\dagger}$, is contained within the set. As the set is infinitely large for continuous data, estimates must be used to find the set.

Predictions can be made by sampling all the possible models within $\mathcal{ILR}(D)$ and creating an interval containing the maximum and minimum values, $\pi(\mathbf{x}) = \left[\underline{\pi(\mathbf{x})}, \overline{\pi(\mathbf{x})}\right]$. When calculating the probability of a value being 1 under the imprecise model, there is an interval probability $\left[\underline{\pi(\mathrm{x})}, \overline{\pi(\mathrm{x})}\right]$, where $\overline{\pi(\mathrm{x})}$ and $\underline{\pi(\mathrm{x})}$ are the maximum and minimum values of $\pi(\mathrm{x})$ calculated from models within the set. As such, when using $\mathcal{ILR}\left(D\right)$ to perform classifications, this interval makes Equation 5.3:

$$\hat{y} = \begin{cases} 1 & \text{if } \underline{\pi}(\mathbf{x}) > C \\ 0 & \text{if } \overline{\pi}(\mathbf{x}) < C \\ [0,1] & \text{if } C \left[\underline{\pi(\mathbf{x})}, \overline{\pi(\mathbf{x})}\right] \end{cases} . \tag{5.6}$$

The final line of this equation returns the *dunno* interval, meaning there is uncertainty in determining whether the datum should be predicted 0 or 1, and the model should therefore abstain from providing a classification. It is left up to the analyst to decide what to do with such a result.

Under such a scenario, there are two approaches to characterising the classifier using a confusion matrix. The first is to consider the intervals directly within the confusion matrix. Calculating statistics derived from the confusion matrix requires careful handling for interval calculations (see Section 2.3). To demonstrate, let $a$ be the number of true positives, $b$ be the number of false positives, $c$ be the number of false negatives, and $d$ be the number of true negatives, where $a$, $b$, $c$ and $d$ are all intervals. Since the number of positive cases is fixed, $a$ and $c$ are oppositely dependent on each other (as $a \to \overline{a}$ then $c \to \underline{c}$), as are $b$ and $d$. These dependencies imply that care must be taken when calculating sensitivity and specificity. Sensitivity needs to be calculated as

$$s = \frac{a}{\underline{a} + \overline{c}}, \tag{5.7}$$

and specificity as

$$t = \frac{d}{\underline{b} + \overline{d}}. \tag{5.8}$$

In this instance, the $s$ and $t$ would themselves both be intervals ($s = [\underline{s}, \overline{s}]$ and $t = \left[\underline{t}, \overline{t}\right]$). $\overline{s}$ and $\overline{t}$ would be the sensitivity and specificity of a system that perfectly classified all the uncertain predictions.

Of course, analysts could use various alternative statistics to describe the classifier's performance, some of which require special care when using imprecise numbers. For instance, *precision* and *recall* are often used within the machine learning literature to assess classifiers. Precision is the fraction of positive predictions that are true positives,

$$\text{precision} = \frac{\text{True Positive}}{\text{Total Number of Positive Predictions}}, \tag{5.9}$$

and recall is analogous to sensitivity. Often quoted alongside these statistics is the $F_1$ score, which is the harmonic mean of these values,

$$F_1 = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{5.10}$$

With intervals, precision needs to be calculated using a single-use expression to ensure that there is no artifactual uncertainty within the calculation,

$$\text{precision} = \frac{1}{1 + \frac{b}{a}}, \tag{5.11}$$

and recall calculated using Equation 5.7.

The $F_1$ score is again best calculated through a single-use expression,

$$F_1 = 2\frac{1}{\frac{TP+b}{a} + 1}, \tag{5.12}$$

where $TP$ is the total number of positive predictions ($TP = \underline{a} + \overline{c}$).

An alternative approach to constructing a confusion matrix with uncertain predictions is to tabulate the dunno predictions in a separate row. If the model returned $u$ true positives, $v$ false positives, $w$ false negatives and $x$ true negatives, but did not make a prediction for $y$ positives and $z$ negatives, then the confusion matrix shown in Table 5.2 can be created.

From this confusion matrix, some useful statistics can be calculated to account for the uncertainty produced by these uncertain classifications. The traditional definitions of sensitivity and specificity can be re-imagined by defining what the *predictive sensitivity* $s'$ as the sensitivity out of the points for which a prediction was made,

$$s' = \frac{u}{u + w}, \tag{5.13}$$

and similarly, the *predictive specificity* $t'$ as the specificity for which a prediction was made,

$$t' = \frac{x}{v + x}. \tag{5.14}$$

Two other statistics are useful to describe the data in Table 5.2. We can define the *positive incertitude* $\sigma$ to be the fraction of positive cases for which the model could not make a prediction,

$$\sigma = \frac{y}{u + w + y}. \tag{5.15}$$

Similarly, the *negative incertitude* $\tau$ can be defined as the total number of negative cases

| | 1 | 0 | Total |
|---|---|---|---|
| Predicted 1 | $u$ | $v$ | $P_+$ |
| Predicted 0 | $w$ | $x$ | $P_-$ |
| No Prediction | $y$ | $z$ | $P_\times$ |
| Total | $T_+$ | $T_-$ | $N$ |

Table 5.2: Alternative confusion matrix where uncertain predictions are tabulated separately.

for which the model could not make a prediction,

$$\tau = \frac{z}{v+x+z}. \tag{5.16}$$

### 5.2.1 Algorithms to Estimate the Set

As the set described in Equation 5.5 is infinitely large, approaches need to be taken to estimate it. Since any analysis of $\mathcal{ILR}(D)$ only requires knowledge of the maximum and minimum $\pi(\mathbf{x})$ values $\left(\left[\underline{\pi(\mathbf{x})}, \overline{\pi(\mathbf{x})}\right]\right)$, the best-possible model would only need to contain the models that produce extreme $\pi(\mathbf{x})$ values.

#### Systematic Search

The ideal method of estimating the set is to search values from within the intervals systematically. This approach requires specifying $P$ as the number of steps within the intervals, then producing a logistic regression model as shown in Algorithm F0. This approach would have complexity $\mathcal{O}(N^P)$ where $P$ is the number of intervals within the set. So, whilst it would be the method most accurate to the best-possible method of computing $\mathcal{ILR}\left(D\right)$ for large datasets, this algorithm would be exceedingly time expensive and thus impractical.

#### Method of Minimum and Maximum Coefficients

Since it is only necessary to find the logistic regression models that correspond to extreme values of $\pi(\mathbf{x})$ for all $\pi(\mathbf{x})$, it is possible to reduce the number of models that need to be contained within $\mathcal{ILR}\left(D\right)$ to only those that make up the envelope of the set. These lines can be estimated as

$$\begin{aligned} \mathcal{ILR}(D) = &\left\{\mathcal{LR}\left(D'_{\underline{\beta_i}}\right), \mathcal{LR}\left(D'_{\overline{\beta_i}}\right)\ \forall i = 0, 1, \ldots, m\right\} \\ &\cup \left\{\mathcal{LR}\left(\underline{D}\right), \mathcal{LR}\left(\overline{D}\right)\right\} \end{aligned}, \tag{5.17}$$

where $D'_{\underline{\beta_0}}$ is the dataset constructed from points within the intervals such that the value of

$\beta_0$ is minimised, $D'_{\overline{\beta_0}}$ is the dataset constructed from points within the intervals such that the value of $\beta_0$ is maximised, and so on. $\underline{D}$ corresponds to the dataset created by taking the lower bound of every interval within $D$, similarly for $\overline{D}$. For a dataset with $m$ features, there are $2m + 2^m$ models that are needed to find the bounds of the set. Algorithm F1 can be used to find this set.

To illustrate, it is useful to first consider a linear regression model with intervals in the dependent variables. Consider the two interval datapoints shown in Figure 5.3, $(x_1, y_1) = ([1, 2], 2)$ and $(x_2, y_2) = ([3, 4], 3)$.

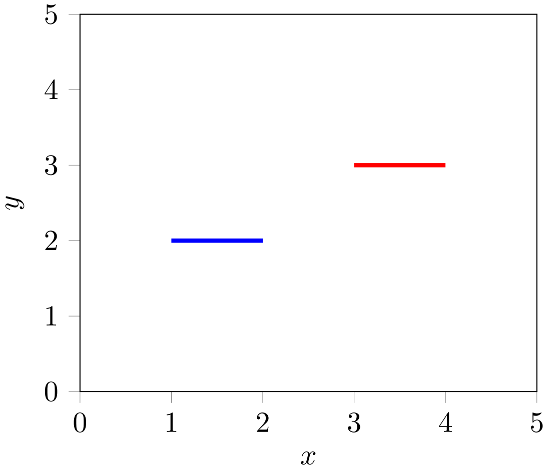


Figure 5.3: Intervals points $(x_1, y_1) = ([1, 2], 2)$ and $(x_2, y_2) = ([3, 4], 3)$.

Whilst there are infinitely many linear regression models ($y = \beta_0 + \beta_1 x$) that could be drawn by selecting points from $x_1$ and $x_2$ there are four extreme values as shown in 5.4. In this instance, the four regression lines (shown in Figure 5.4) are:

$$y = x \tag{5.18a}$$

$$y = \frac{1}{3}x + \frac{5}{3} \tag{5.18b}$$

$$y = x - 2 \tag{5.18c}$$

$$y = \frac{1}{2}x + \frac{3}{2} \tag{5.18d}$$

If we consider these four lines as a set and consider the imprecise regression model as the envelope of this set, then we get the band shown in Figure 5.5. Six segments make up this band ($AB$, $BC$, $CD$, $EF$, $FG$, $GH$). These lines correspond to lines with the minimum and maximum $\beta_0$ and $\beta_1$ values plus the line drawn with the left-most and right-most values from within the intervals.

**Algorithm F0:** Systematic search to find $\mathcal{ILR}\left(D\right)$ if $D$ has interval uncertainty within its features.

```
Input: $D = \left\{ \left( \left( \left[ \underline{x_j^{(i)}}, \overline{x_j^{(i)}} \right] \quad \forall i = 0, \ldots, m \right), y_j \right) \quad \forall j = 0, \ldots, n \right\}$, $Q$ steps
$\mathcal{ILR}\left(D\right) \leftarrow \{\}$;
$\Delta \leftarrow \left\{ \frac{i-1}{Q-1} \ \forall i = 1, \ldots, Q \right\}$;
for all combinations $\left\{ \delta_j^{(i)} \right\} \in \Delta$ do
    $D' \leftarrow D$;
    for all $i \in \{0, \ldots, m\}$ do
        for all $j \in \{0, \ldots, n\}$ do
            $D_j'^{(i)} \leftarrow \left\{ \underline{x_j^{(i)}} + \delta_j^{(i)} \left( \overline{x_j^{(i)}} - \underline{x_j^{(i)}} \right) \right\}$
        end
    end
    $\mathcal{ILR}\left(D\right) \leftarrow \mathcal{ILR}\left(D\right) \cup \left\{ \mathcal{LR}\left(D'\right) \right\}$;
end
Output: $\mathcal{ILR}\left(D\right)$
```

**Algorithm F1:** Method to find $\mathcal{ILR}\left(D\right)$ if $D$ has interval uncertainty within its features by calculating the minimum and maximum intercept and coefficients.

```
Input: $D = \left\{ \left( \left( \left[ \underline{x_j^{(i)}}, \overline{x_j^{(i)}} \right] \quad \forall i = 0, \ldots, m \right), y_j \right) \quad \forall j = 0, \ldots, n \right\}$
$\underline{D} = \left\{ \left( \left( \underline{x_j^{(i)}} \ \forall i = 0, \ldots, m \right), y_j \right) \quad \forall j = 0, \ldots, n \right\}$;
$\overline{D} = \left\{ \left( \left( \overline{x_j^{(i)}} \ \forall i = 0, \ldots, m \right), y_j \right) \quad \forall j = 0, \ldots, n \right\}$;
$\mathcal{ILR}\left(D\right) \leftarrow \left\{ \mathcal{LR}\left(\underline{D}\right), \mathcal{LR}\left(\overline{D}\right) \right\}$;
for $i \leftarrow 0$ to $m$ do
    Using stochastic optimisation find $D'_{\underline{\beta_i}}$ such that $\mathcal{LR}\left(D'_{\underline{\beta_i}}\right)$ has the minimum
      value of $\beta_i$;
    $\mathcal{ILR}\left(D\right) \leftarrow \mathcal{ILR}\left(D\right) \cup \left\{ \mathcal{LR}\left(D'\right) \right\}$;
    Using stochastic optimisation find $D'_{\overline{\beta_i}}$ such that $\mathcal{LR}\left(D'_{\overline{\beta_i}}\right)$ has the maximum
      value of $\beta_i$;
    $\mathcal{ILR}\left(D\right) \leftarrow \mathcal{ILR}\left(D\right) \cup \left\{ \mathcal{LR}\left(D'\right) \right\}$;
end
Output: $\mathcal{ILR}\left(D\right)$
```

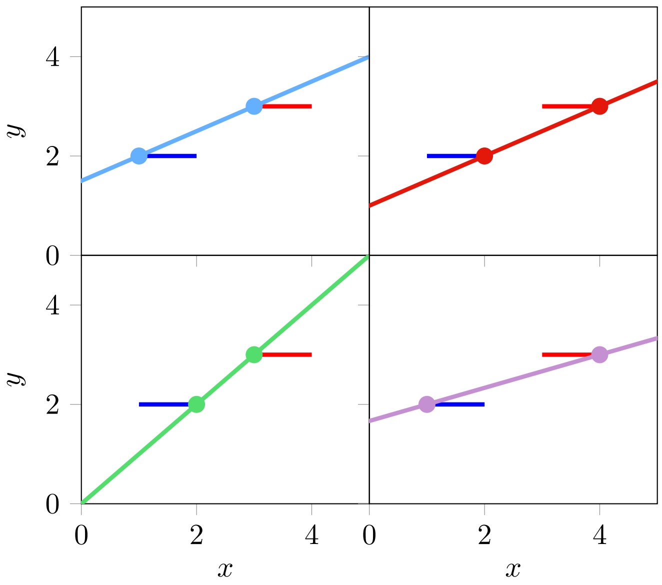


Figure 5.4: Four extreme regression models that can be fitted from the two intervals.

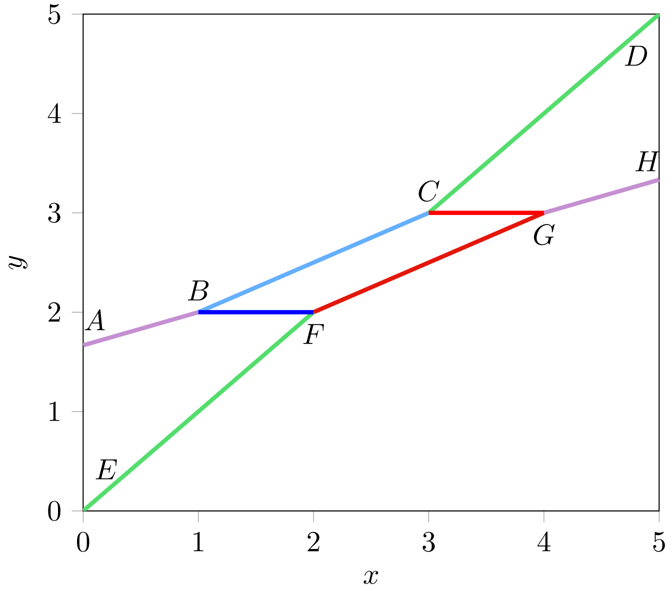


Figure 5.5: Envelope of the extreme lines shown in Figure 5.4.

In theory, there are six possible models; it just so happens to be the case that the line which contains $\underline{\beta_0}$ also has $\overline{\beta_1}$ and vice versa. Figure 5.6 shows 100 regression models fitted using 100 Monte Carlo samples for values within the intervals. The whole band between the black lines would have been filled with enough samples or systematic sampling.

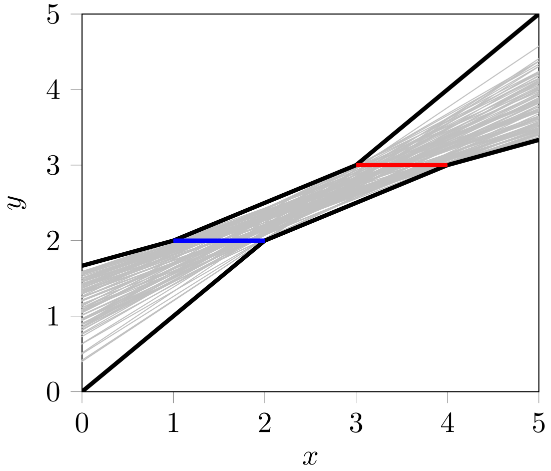


Figure 5.6: One hundred regression models (shown in grey) fitted on Monte Carlo samples for values within the intervals.

If we have four intervals instead of two, we can test how well the method finds the minimum and maximum coefficient values alongside the all-left bound and all-right bound models. Figure 5.7 shows four intervals and the band bounded by the six lines suggested to be plotted. As before, we can sample values from within the intervals to check the performance of the band.

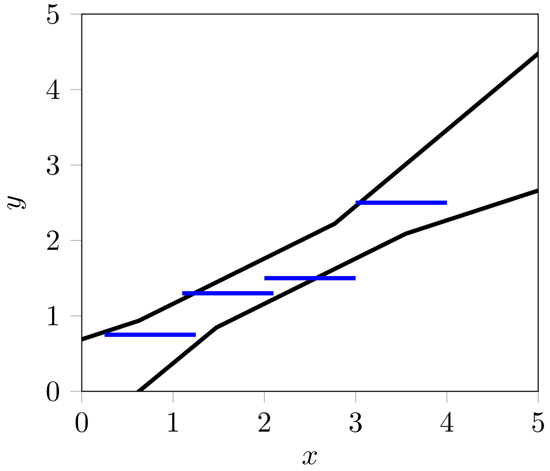


Figure 5.7: Imprecise linear regression model for the 4 intervals shown.

Again we can test whether this imprecise model bounds all possible regression models for values within the intervals. If we do a systematic sample of points within the four intervals, we can find all valid linear regression models consistent with the intervals. The envelope of these models is shown with the grey band in Figure 5.8. As can be seen, this band is not entirely within the black lines. To measure how good an of estimate the method is, we can let $A$ be the area of the systematic bounds that are outside the bounds produced by the proposed algorithm as a fraction of the total area found systematically. Ideally, we would see $A = 0$, implying that the set perfectly covered the systematic bounds. For the model shown in Figure 5.8, doing this gives a value $A = 0.0311$.

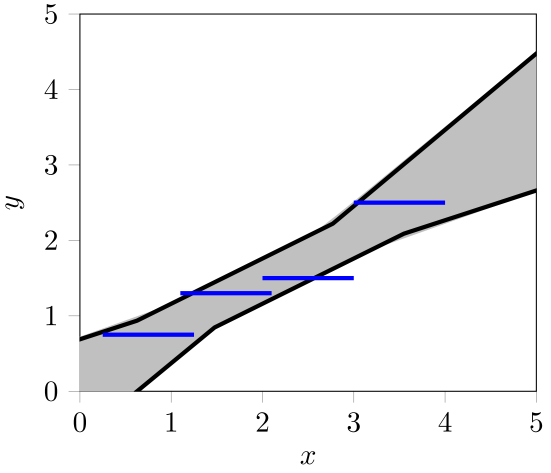


Figure 5.8: Comparison between the imprecise linear regression and bounds produced by a systematic search.

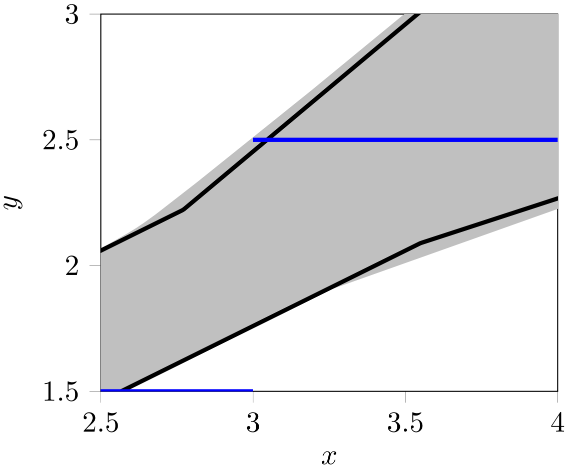


Figure 5.8a: Zoomed in version of Figure 5.8.

Moving to logistic regression, Equation 5.1 can be written as

$$\pi(x) = \frac{1}{1 + \exp -r(x)} = \sigma(r(x)) \tag{5.19}$$

where $r(x) = \beta_0 + \beta_1 x$ and $\sigma$ is the sigmoid function. Since we know that the 6-line method is suitable for estimating the envelope of $r(x)$ and the sigmoid function is monotonic, we can use this method for logistic regression. The is the set represented by Equation 5.5 produced by Algorithm F1.

If we consider two intervals with opposite binary labels, $(x_1, y_1) = ([1, 2], 0)$ and $(x_2, y_2) = ([3, 4], 1)$, shown in Figure 5.9, we can see that the endpoints of the intervals will produce the minimum and maximum logistic regression models, as shown in Figure 5.10. As before, these lines represent the minimum and maximum coefficient values.

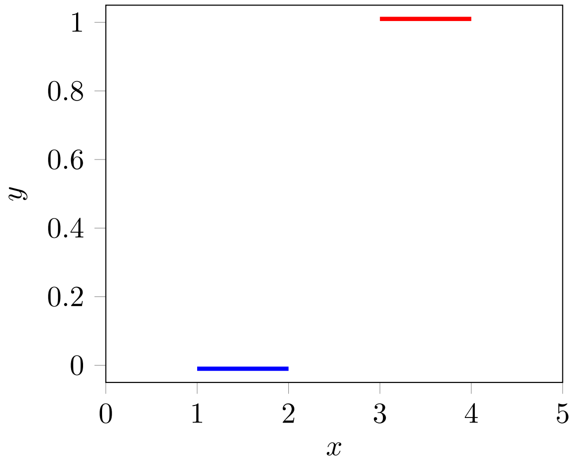


Figure 5.9: Two intervals with binary labels

Again we can test whether the envelope of these four models (in this case, just the two lines that represent the models fitted on the endpoints of the intervals) will contain all possible models for values within the intervals via systematic sampling. This is shown in Figure 5.11. Figure 5.12 shows all six lines for the intervals shown compared to the systematic method, with the black dashed lines representing the envelope of the set, with $A = 0.0160$.

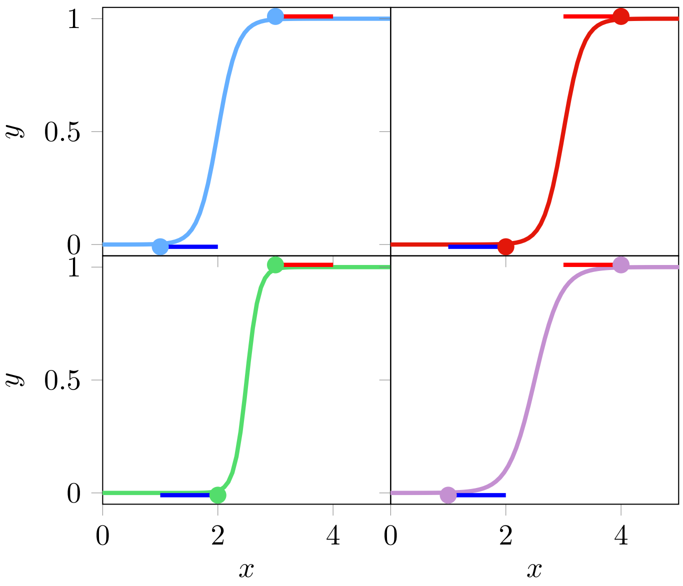


Figure 5.10: 4 extreme regression models that can be fitted from the two intervals.

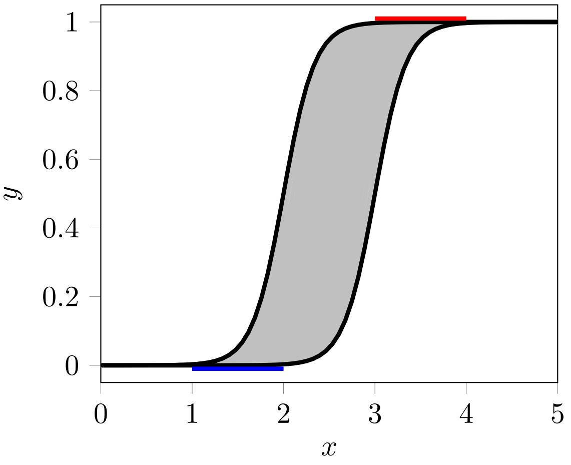


Figure 5.11: Envelope of the logistic regression models shown in Figure 5.10 and models produced by systematically sampling the intervals (grey).

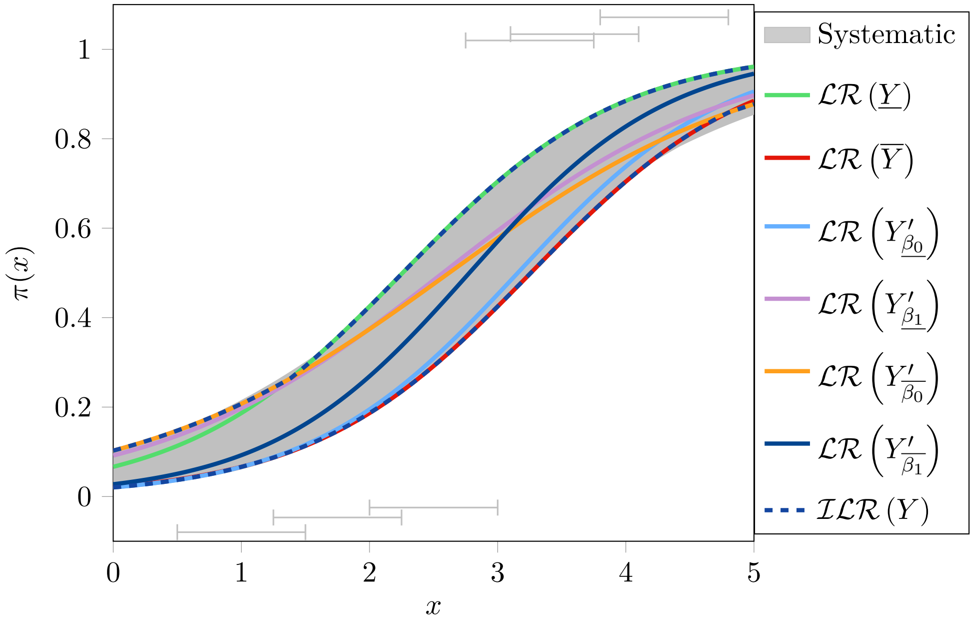


Figure 5.12: Comparison of systematic approach against Algorithm F1.

### Minimum and Maximum Spread Approach

If the dataset contains a large number of intervals, then due to the increasing complexity of the optimisation, Algorithm F1 may take a prohibitively long time to compute. In such a situation, Algorithm F2 can be used to find the imprecise model. This algorithm uses the heuristic that the extreme bounds are likely to be associated with the minimum and maximum spread of points around particular values. The complexity of this approach is $\mathcal{O}(2(1+P))$.

**Algorithm F2:** Minimum and Maximum spread approach to estimate $\mathcal{ILR}\left(D\right)$ if $D$ has interval uncertainty within its features. The procedures for finding the minimum and maximum spread of points around a particular value can be found in Procedures 5.1 and 5.2.

```
Input: $D = \left\{\left(\left(\left[\underline{x_j^{(i)}}, \overline{x_j^{(i)}}\right] \quad \forall i = 0,\ldots,m\right), y_j\right) \quad \forall j = 0,\ldots,n\right\}$
$\underline{D} = \left\{\left(\left(\underline{x_j^{(i)}} \ \forall i = 0,\ldots,m\right), y_j\right) \quad \forall j = 0,\ldots,n\right\}$;
$\overline{D} = \left\{\left(\left(\overline{x_j^{(i)}} \ \forall i = 0,\ldots,m\right), y_j\right) \quad \forall j = 0,\ldots,n\right\}$;
$\mathcal{ILR}\left(D\right) \leftarrow \left\{\mathcal{LR}\left(\underline{D}\right), \mathcal{LR}\left(\overline{D}\right)\right\}$;
for all $k = 1,\ldots,P$ do
    Find $\underline{T}$ such that $\pi_{\mathcal{LR}(\underline{D})}(T) = \frac{k}{P+1}$;
    Find $\overline{T}$ such that $\pi_{\mathcal{LR}(\overline{D})}(T) = \frac{k}{P+1}$;
    $d_{min} \leftarrow$ minimum spread of points around $\underline{T}$;
    $d_{max} \leftarrow$ maximum spread of points around $\overline{T}$;
    $\mathcal{ILR}\left(D\right) \leftarrow \mathcal{ILR}\left(D\right) \cup \left\{\mathcal{LR}\left(d_{min}\right), \mathcal{LR}\left(d_{max}\right)\right\}$;
end
Output: $\mathcal{ILR}\left(D\right)$
```

To illustrate this method, we need to reconsider Figure 5.10. Instead of considering that the four lines correspond to the minimum and maximum intercept and coefficient values, note that, as shown in Figure 5.13, the lines produced also represent the minimum and maximum spread of values around $x = 1$, $x = 2.5$ and $x = 4$.

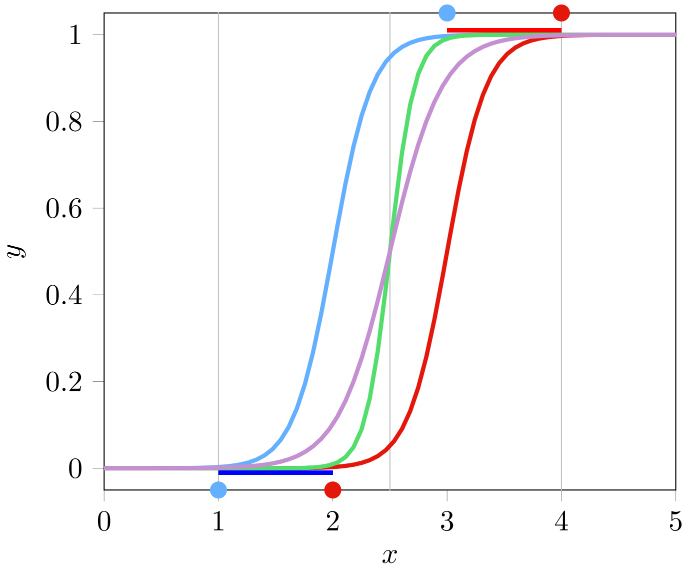


Figure 5.13: Repoduction of Figure 5.10 showing that the extreme models can be considered as the lines that correspond to the minimum and maximum spread of values around $x = 1$, $x = 2.5$ and $x = 4$ (shown with grey lines).

This leads to another way of approximating Equation 5.5. If we $x = 1$ and $x = 4$ are the minimum and maximum values from within the dataset, the minimum and maximum spread around these points represent the all-left and all-right bound models. The $x = 2.5$ corresponds to the minimum and maximum slope of the logistic regression curve. Finding the $x$ value that corresponds to this extreme case is trivial when there are only two intervals, however, is challenging for more complicated datasets.

One approach is to select $P$ points from within the range of the dataset and the minimum and maximum values themselves, giving $2 + P$ values to sample. For a dataset with $m$ features, this leads to $(2 + P)^m$ models that would need to be found.

An alternative approach is to first fit the all-left and all-right models ($\mathcal{LR}\left(\underline{D}\right)$ and $\mathcal{LR}\left(\overline{D}\right)$), then select $P$ $\pi$ values ($\pi = \left\{\frac{p}{P+1} \ \forall p = 1, \ldots, P\right\}$). For each of these $P$ values we can find $\underline{T}$, such that $\pi_{\mathcal{LR}(\underline{D})}(T) = P$ and $\overline{T}$, such that $\pi_{\mathcal{LR}\left(\overline{D}\right)}(T) = P$. We can then find the datasets that correspond to the minimum and maximum spread of values using Algorithm 5.1 and 5.2 respectively. This approach has complexity $2(1 + P)$ and is the approach described within Algorithm F2.

The bounds produced by this method for the same six intervals from Figure 5.12 are shown in Figure 5.14. This plot has $A = 0.0351$ compared to the systematic search method.

**Algorithm 5.1:** Procedure to find the minimum spread of the intervals in $D$ around $T$.

```
Data: $D$, $T$
for all Intervals ($I = [\underline{i}, \overline{i}]$) in $D$ do
    if $T \in P$ then
        $D_{min}(I) = T$;
    else if $T < P$ then
        $D_{min}(I) = \underline{i}$;
    else
        $D_{min}(I) = \overline{i}$;
Output: $D_{min}$
```

**Algorithm 5.2:** Procedure to find the maximum spread of the intervals in $D$ around $T$.

```
Data: $D$, $T$
for all Intervals ($I = [\underline{i}, \overline{i}]$) in $D$ do
    if $T \in P$ then
        if $|\underline{i} - T| > |\overline{i} - T|$ then
            $D_{max}(I) = \underline{i}$;
        else
            $D_{max}(I) = \overline{i}$
    else if $T < P$ then
        $D_{max}(I) = \overline{i}$;
    else
        $D_{max}(I) = \underline{i}$;
Output: $D_{max}$
```

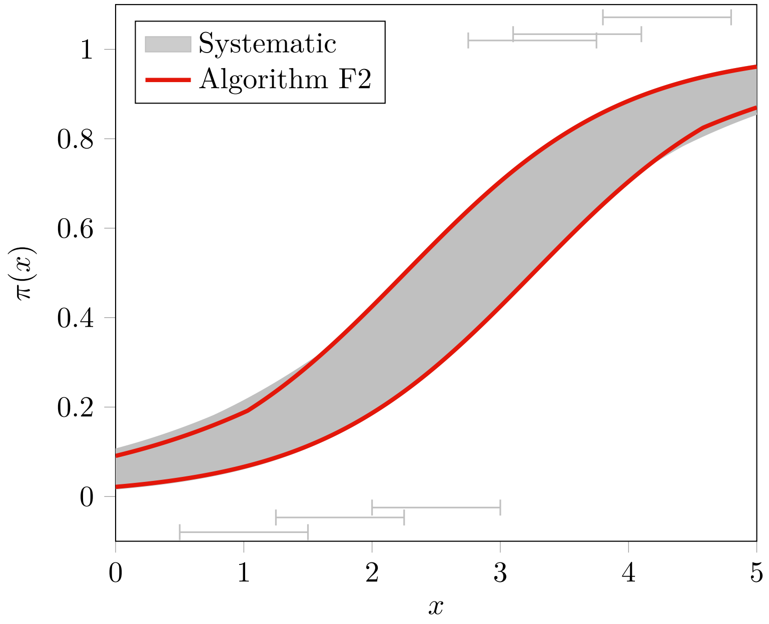


Figure 5.14: Comparison of systematic approach against Algorithm F2 with $P = 100$.

### Monte Carlo

A straightforward and computationally expedient way of estimating the set is to use Monte Carlo to find valid models. This approach, as shown in Algorithm F3, requires the analyst to specify a desired number of iterations, then, for every iteration, make a new dataset by sampling a random value from all intervals within the dataset, and then fit a model on this new dataset. Often this sampling assumes the equidistribution hypothesis and thus models the intervals as a uniform distribution. This approach is likely to underestimate the bounds as Monte Carlo methods are unlikely to find the extreme models [168].

Figure 5.15 shows a comparison between a Monte Carlo search with $10^6$ iterations and a systematic search for the six intervals shown. As expected, the bounds from the Monte Carlo method fail to enclose the total area revealed by the systematic search. In this case, $A = 0.166$. A higher number of iterations would have led to a smaller $A$ value.

## 5.2.2 Alternative Methods

Several authors have suggested different approaches to compute logistic regression models with interval uncertainty. This section will consider three methods to compare with the methods presented in this paper.

### Midpoint

The most straightforward approach to dealing with interval data is to produce a precise dataset by replacing the intervals with their midpoints and then fitting a dataset with this midpoint data. i.e.

$$D_m = \left\{ \left( \left( \frac{\underline{x_j^{(i)}} + \overline{x_j^{(i)}}}{2} \right), y_j \right) \;\; \forall \left[ \underline{x_j^{(i)}}, \overline{x_j^{(i)}} \right] \in D \right\}, \tag{5.20}$$

then

$$\mathcal{LR}_m\left(D\right) = \mathcal{LR}\left(D_m\right). \tag{5.21}$$

This dataset can then be used as a precise logistic regression model described in Section 5.1.

### de Souza

de Souza et al. [244, 245] introduce several methods for characterising the uncertainty with interval features. They conclude that the best method is to perform two separate logistic regressions on the lower and upper bounds of the intervals and average the posterior

**Algorithm F3:** Monte Carlo search to find $\mathcal{ILR}\left(D\right)$ if $D$ has interval uncertainty within its features.

```
Input: $D = \left\{ \left( \left( \left[ \underline{x_j^{(i)}}, \overline{x_j^{(i)}} \right] \quad \forall i = 0, \ldots, m \right), y_j \right) \quad \forall j = 0, \ldots, n \right\}$, $P$ steps
$\mathcal{ILR}\left(D\right) \leftarrow \{\}$;
for $P$ iterations do
    $D' \leftarrow D$;
    $D_j'^{(i)} \leftarrow$ random value in $\left[ \underline{x_j^{(i)}}, \overline{x_j^{(i)}} \right]$;
    $\mathcal{ILR}\left(D\right) \leftarrow \mathcal{ILR}\left(D\right) \cup \left\{ \mathcal{LR}\left(D'\right) \right\}$;
end
Output: $\mathcal{ILR}\left(D\right)$
```

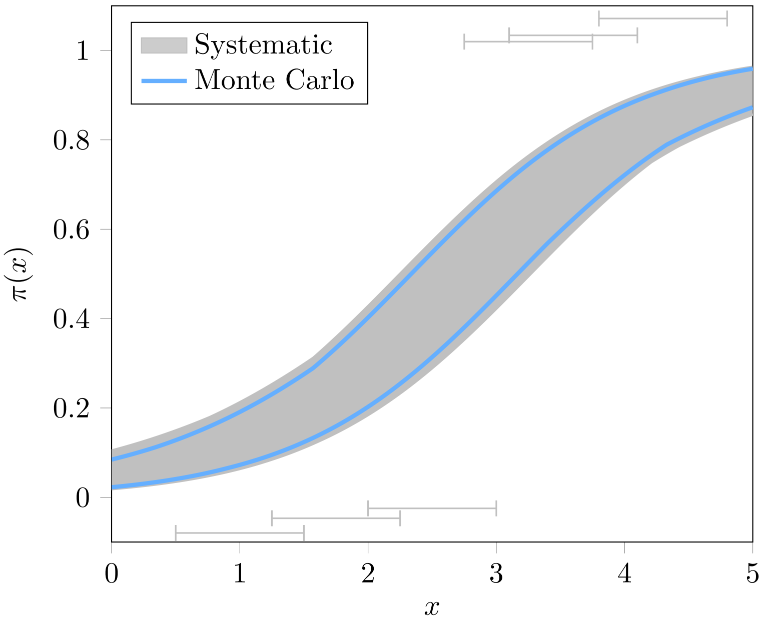


Figure 5.15: Comparison of systematic approach against a Monte Carlo sample with $10^6$ iterations.

probabilities to obtain a pooled posterior probability. They find

$$\mathcal{LR}_{dS}\left(D\right) = \left\{\mathcal{LR}\left(\underline{D}\right), \mathcal{LR}\left(\overline{D}\right)\right\}, \tag{5.22}$$

and then reduce this to a single logistic regression model based on the average of the outputted probabilities:

$$\pi_{\mathcal{LR}_{dS}(D)}(\mathbf{x}) = \frac{\pi_{\mathcal{LR}(\underline{D})}(\mathbf{x}) + \pi_{\mathcal{LR}\left(\overline{D}\right)}(\mathbf{x})}{2}. \tag{5.23}$$

### Billard-Diday

Billard and Diday [241] propose a method, based upon Bertrand [240], for characterising interval uncertainties within linear regression that they suggest can be easily extended to logistic regression. Their method assumes that each value from within the interval is equally likely, and therefore constructs the logistic regression models as the uniform mixture of $N$ logistic regression models that are fitted from random samples,

$$\mathcal{LR}_{BD}\left(D\right) = \left\{\mathcal{LR}\left(D_k\right) : D_k = \left\{\left\{\left((r_j^{(i)}), y_j\right)\right\} \; r_j'^{(i)} \in \left[\underline{x_j^{(i)}}, \overline{x_j^{(i)}}\right]\right\} \; k = 0, \ldots, N\right\}. \tag{5.24}$$

Like de Souza, they then average the probability from all models,

$$\pi_{\mathcal{LR}_{BD}(D)}(\mathbf{x}) = \frac{1}{N} \sum_{\forall l \in \mathcal{LR}_{BD}(D)} \pi_l(\mathbf{x}). \tag{5.25}$$

This method is computationally the same as Algorithm F3 but takes the average of the found models to produce a precise final model instead of taking the envelope to produce an imprecise model.

### 5.2.3 Comparison of Methods

Dataset $D_1$ from 5.1 has been intervalised into dataset $D_2$ using the following transformation:

$$x_i' = [m - \epsilon, m + \epsilon]$$

where $m$ is a number drawn from the triangular distribution $\mathrm{T}(x_i - \epsilon, x_i + \frac{\epsilon}{6}, x_i + \epsilon)$ with $\epsilon = 0.375$ for all $x_i \in X$. With this dataset we can use Algorithms F1, F2 and F3 to construct $\mathcal{ILR}\left(D_2\right)$, as is shown in Figure 5.16. It would have been too computationally expensive to perform a systematic search using Algorithm F0.

Figure 5.16 shows that Algorithm F1 produces the most comprehensive bounds and is therefore most likely to represent the entire set described by Equation 5.5. The Monte Carlo approach in Algorithm F3 produces the narrowest bounds with Algorithm F2.

For comparison, logistic regression models have been fitted using the methods described in Section 5.2.2; these models are shown in Figure 5.17. Whilst there are subtle differences between the different approaches, it is clear that they are all approximately equal. This equivelence is unsurprising as they all implicitly make the equidistribution assumption that a uniform distribution can represent the intervals.

It is also possible to consider situations where data is intervalised differently. Figure 5.18 shows four different intervalisations of dataset $X$ as described in Table 5.3. In plot (a), the intervalisation has occurred by placing the true value at the left edge of the interval; similarly, in (b), the true value is at the right edge of the interval. In (c), the value of $x$ impacts the interval's width, and in (d), the label impacts the intervalisation. In this figure, imprecise models have been fitted on the datasets using Algorithms F1 F2 and F3 alongside the 'true' model (from Figure 5.1) and the midpoint model. The de Souza and Billard-Diday methods are not shown due to their similarity with the midpoint model.

Looking at all these figures, we see that the imprecise model produced by Algorithms F1 and F2 always bounds the base model. As a result, any interval regression analysis performed would be guaranteed to bound the true model. The model produced by Algorithm F3 fails to do so in (a), (b) and (d). In (a) and (b) the fact that $\mathcal{ILR}_{\mathrm{F1}}(D_{2\circ}) \nsubseteq \mathcal{ILR}_{\mathrm{F2}}(D_{2\circ})$ and $\mathcal{ILR}_{\mathrm{F2}}(D_{2\circ}) \nsubseteq \mathcal{ILR}_{\mathrm{F1}}(D_{2\circ})$ suggests neither Algorithm F1 nor Algorithm F2 is superior to the other.

The figure also shows that there can be significant differences between the base and midpoint models. The alternative approaches provide a good approximation of the true value if the equidistribution hypothesis can be justified, as in plot (c). If the intervalisation depends on the outcome, then the approaches appear inadequate, as is shown in plot (d). This implies that the alternative approaches, and Algorithm F3, should only be used in cases within which one can assume that the data has been intervalised independent of either the true underlying value or the outcome status and each value within the interval is equally likely.

In many real-world datasets, the assumptions that the alternative methods rely on may be unjustified or untennable, and in those scenarios, only the complete imprecise method would guarantee coverage of the true model.

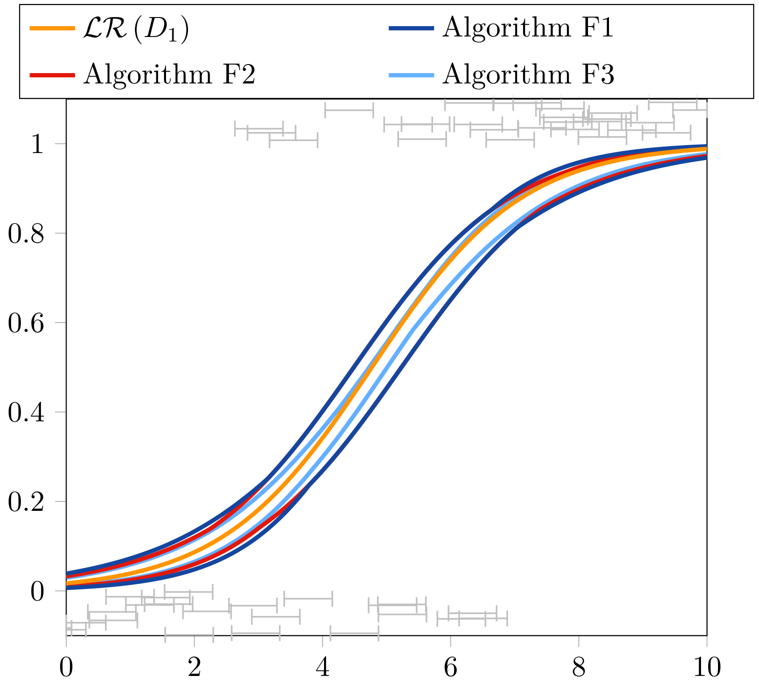


Figure 5.16: Imprecise logistic regression models fitted using Algorithms F3 (with $10^4$ iterations), F1 and F2 for the interval data (jittered for clarity). $\mathcal{LR}\,(D_1)$ represents the 'true' model from Figure 5.1.

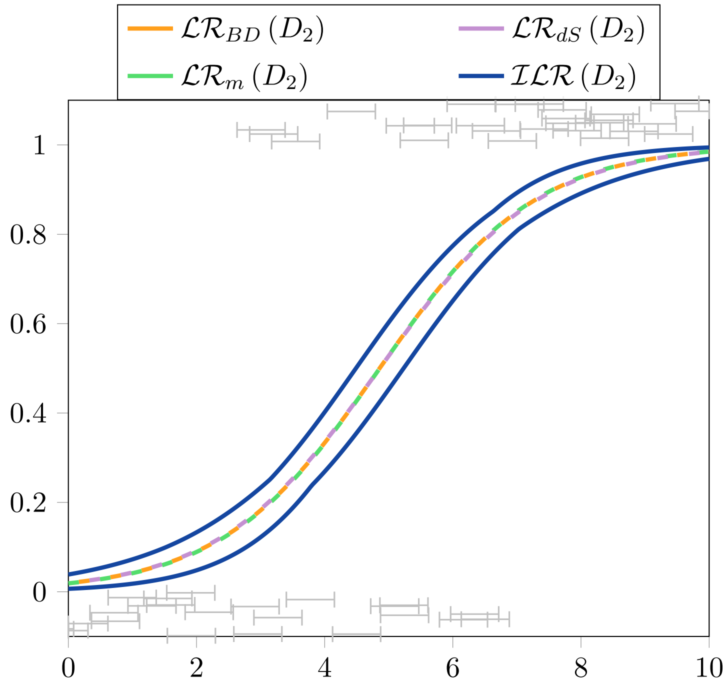


Figure 5.17: Imprecise logistic regression model fitted using Algorithm F1 and logistic regression models fitted using the alternative methods in Section 5.2.2 for the interval dataset (jittered for clarity).

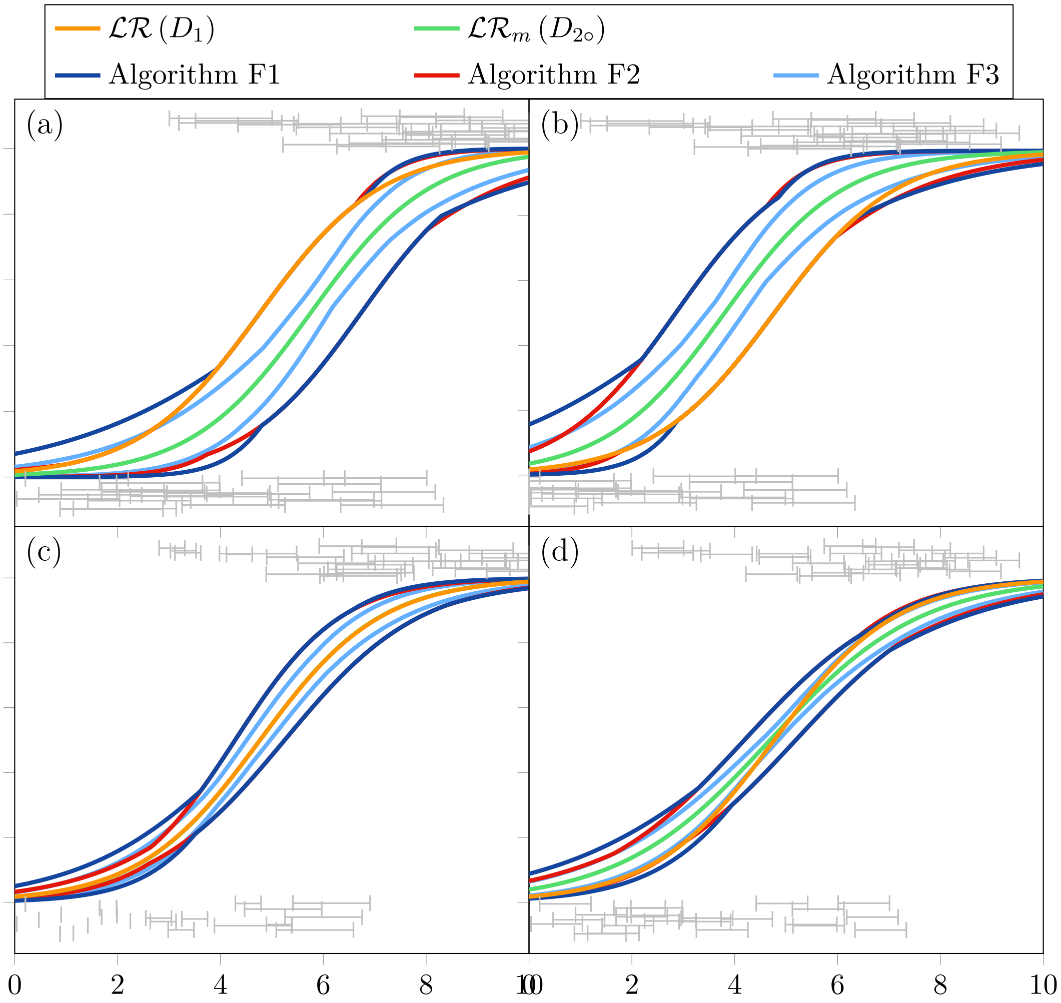


Figure 5.18: Logistic Regression models fitted on interval data that has been intervalised in the biased ways shown in Table 5.3. In all plots, $\mathcal{LR}(D_1)$ is the precise model shown in Figure 5.1.

| Plot | Intervalisation |
|---|---|
| (a) | $D_{2a} = \{([x_i, x_i + 2], y_i)\ \ \forall (x_i, y_i) \in X\}$ |
| (b) | $D_{2b} = \{([x_i - 2, x_i], y_i)\ \ \forall (x_i, y_i) \in X\}$ |
| (c) | $D_{2c} = \Bigg\{\Bigg(\Bigg(\begin{cases} x_i \text{ if } x_i < 2.5 \\ [m - 0.25, m + 0.25],\ \ m \in U(x_i \pm 0.25) \text{ if } 2.5 \leq x_i < 5.0 \\ [m - 0.75, m + 0.75],\ \ m \in U(x_i \pm 0.75) \text{ if } 5.0 \leq x_i < 7.5 \\ [x_i, x_i + 1.5],\ \text{ if } 7.5 < x_i \end{cases}, y_i\Bigg) \forall (x_i, y_i) \in X\Bigg\}$ |
| (d) | $D_{2d} = \Bigg\{\Bigg(\Bigg(\begin{cases} [x_i, x_i + 1.5] \text{ if } y_i = 1 \\ [x_i - 1.5, x_i] \text{ otherwise} \end{cases}, y_i\Bigg) \forall (x_i, y_i) \in X\Bigg\}$ |

Table 5.3: Intervalisations used in Figure 5.18.

### 5.2.4 Red Wine Example

In order to demonstrate the methodology on a real-world dataset, we can use the red wine dataset from [262]. This dataset contains 11 covariates* that can be used to predict the quality of the wine sample based on a scale from 0 to 10. In order to provide a binary classification, we define wine as 'good' if it has quality $\geq 6$. The dataset contains 1599 samples, of which 855 have been classified as good wine. Let $R$ be the dataset with added uncertainties and $R^\dagger$ be the original (true) dataset.

In order to fit the logistic regression model, the dataset has been split into training and test subsets containing half the data in each. To intervalise the data, values have been turned into intervals based on the number of significant figures within the model (for example, 0.5 would become $[0.45, 0.55]$). An imprecise model can then be fitted on the dataset using Algorithm F2.

It is helpful to consider visualisations when discussing the classifier's performance (Figure 5.19). The simplest of these is the scatter plot shown in Figure 5.19a. From the plot, most of the wines rated as good were given a high $\pi$ value, and vice versa for the bad wines–although no wine was given a very low $\pi$. There is, however, substantial overlap between the two groups. The plot also shows that some wines have a wide interval $\pi$. The plot also shows the size of the intervals for $\pi$. In this instance, all intervals are reasonably consistent and not overly wide.

We can also construct ROC plots and calculate their AUCs, as shown in Figure 5.19b. In this plot, we can see that the $\mathcal{LR}\left(R^\dagger\right)$ and $\mathcal{LR}_m\left(R\right)$ curves are only subtly distinguishable from each other. The same is also true of $\mathcal{LR}_{dS}\left(R\right)$ and $\mathcal{LR}_{BD}\left(R\right)$. The imprecise model bounds all of these. Additionally, for the imprecise model, we can plot a ROC curve for when the model abstains from making a prediction. In this instance $s'$ is plotted against $fpr'(= 1 - t')$. This ROC curve outperforms the others.

The AUC for the curves are:

$$AUC\left[\mathcal{ILR}\left(R\right)\right] = [0.742, 0.872],$$

$$AUC\left[\mathcal{ILR}\left(R\right)(\mathit{Abstain})\right] = 0.834,$$

$$AUC[\mathcal{LR}\left(R^\dagger\right)] \approx AUC\left[\mathcal{LR}_m\left(R\right)\right] = 0.818,$$

and

$$AUC[\mathcal{LR}_{dS}\left(R\right)] \approx AUC[\mathcal{LR}_{BD}\left(R\right)] = 0.813.$$

*Fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol.

| | Good Wine | Bad Wine | Total |
|---|---|---|---|
| Predicted Good | [279,329] | [68,109] | [347,438] |
| Predicted Bad | [99,149] | [263,304] | [362,453] |
| Total | 428 | 372 | 800 |

(a) Tabulating inconclusive results as [0,1] intervals.

| | Good Wine | Bad Wine | Total |
|---|---|---|---|
| Predicted Good | 279 | 68 | 347 |
| Predicted Bad | 99 | 263 | 362 |
| No Prediction | 50 | 41 | 91 |
| Total | 428 | 372 | 800 |

(b) Tabulating inconclusive results separately.

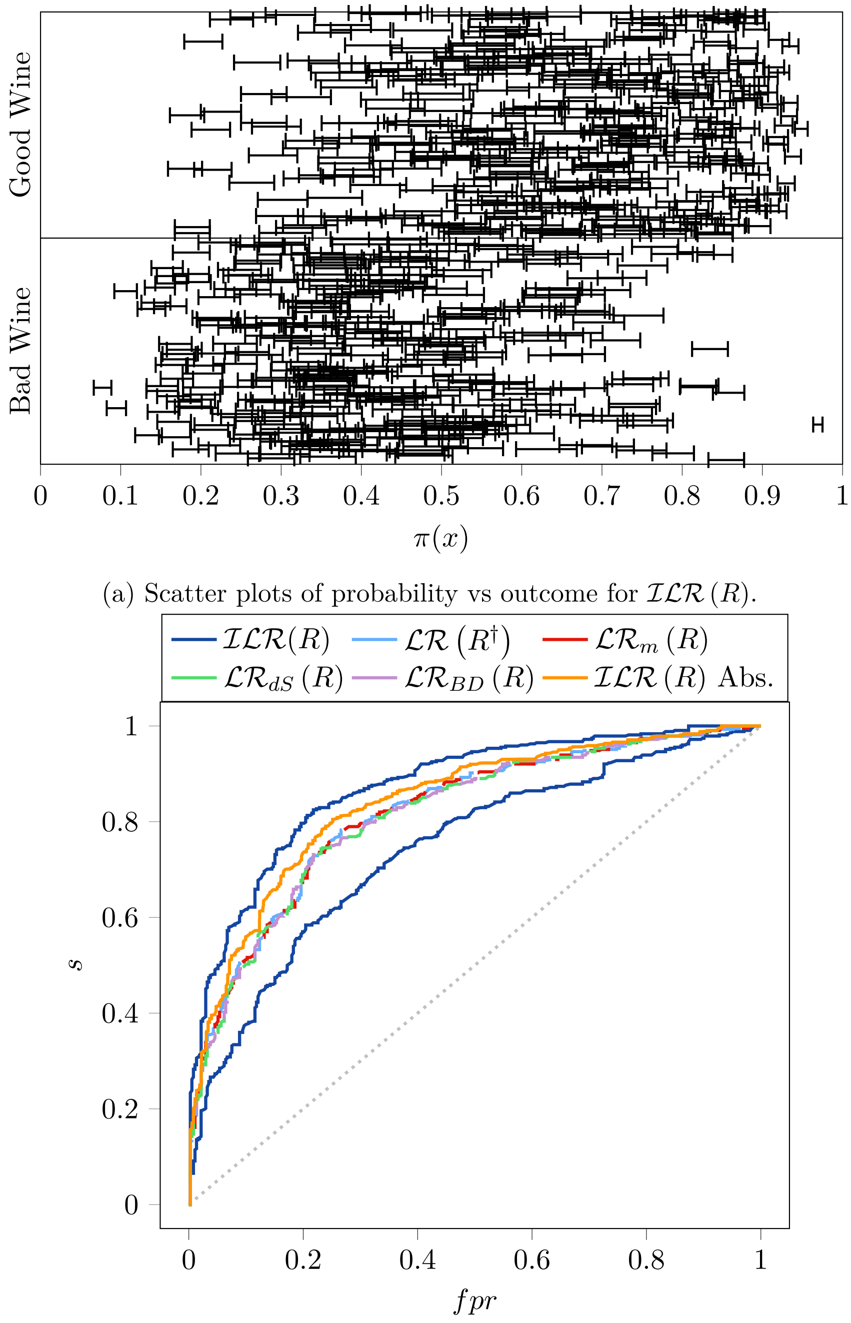
Table 5.4: Two possible confusion matrices for 100 test samples from the imprecise logistic regression model shown in Figure 5.12.

Classifications about whether a wine is good or not can be made. Selecting a threshold value of 0.5 gives the confusion matrices shown in Table 5.4. There are two possible interpretations of how the intervals can be expressed within the confusion matrices shown in Table 5.4a. The intervals are plotted directly within the matrix giving the following statistics: $s = [0.652, 0.769]$ and $t = [0.707, 0.817]$. Allowing the model to abstain as in Table 5.4b implies that there are 91 wines for which a prediction could not be made solely as a result of the imprecision of the model. The summary statistics from this confusion matrix are: $s' = 0.738$, $t' = 0.795$, $\sigma = 0.117$ and $\tau = 0.110$.

## 5.3 Uncertainty in Labels

This set-based approach can be extended to the situation where there is uncertainty about the outcome status meaning there are some points for which we do not know the binary classification and can be represented as the dunno interval $[0, 1]$. In this situation the dataset $D$ contains $p$ variables with corresponding labels $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_p, y_p)$ but also $q$ variables for which the label is unknown $(\mathbf{x}_{p+1},\ ), (\mathbf{x}_{p+2},\ ), \cdots, (\mathbf{x}_{p+q},\ )$. For simplicity, we shall refer to the points with labels as being in set $d$ and those without labels in set $u$, $D = d \cup u$.

It is important to note that, whilst formally, all possible unobserved – and therefore unlabelled – $\mathbf{x}$ values could be considered to be in set $u$, this should not be considered the case. The datapoints that are in $u$ only contain $\mathbf{x}$ values that have been observed, but for which the label is unidentified for some reason. This may be due to a participant dropping out of a medical trial before it concludes or there if there are disagreement between expert labellers



(a) Scatter plots of probability vs outcome for $\mathcal{ILR}\left(R\right)$.

(b) Receiver operating characteristic curve for the simple example with added uncertain classifications.

Figure 5.19: Plots to show the discriminatory performance of the logistic regression models for the red wine example.

about the true clasification. Thus, it is not the case that one can increase the number of points in $u$ by introducing arbitrary new $\mathbf{x}$ values.

Traditional analysis may ignore all the points in $u$. However, they can be included within the analysis by considering the set of logistic regression models trained on all possible datasets that are valid based upon the uncertainty. This set of datasets can be created by setting all unlabelled values to 0, then setting all unlabelled values to 1, and all combinations thereof, i.e.

$$\mathcal{ILR} = \left\{ \mathcal{LR}(D') \; \forall D' \in \left\{ \begin{array}{c} d \cup \{(\mathbf{x}_{p+1}, 0), \cdots, (\mathbf{x}_{p+q}, 0)\} \\ d \cup \{(\mathbf{x}_{p+1}, 0), \cdots, (\mathbf{x}_{p+q}, 1)\} \\ \vdots \\ d \cup \{(\mathbf{x}_{p+1}, 1), \cdots, (\mathbf{x}_{p+q}, 0)\} \\ d \cup \{(\mathbf{x}_{p+1}, 1), \cdots, (\mathbf{x}_{p+q}, 1)\} \end{array} \right\} \right\}. \tag{5.26}$$

There are $2^q$ possible logistic regression models within this set. An imprecise logistic regression model can then be created by finding the envelope of the set, as shown in Algorithm L1. As the computational time for this algorithm increases as $\mathcal{O}(2^q)$, then as $q$ increases finding the bounds by calculating the envelope for all possible combinations can become computationally expensive.

Algorithm L2 reduces the number of models that need to be fitted to find an estimate for the imprecise bounds. This algorithm first finds the logistic regression model assuming all uncertain labels are 0 and the logistic regression model assuming all uncertain labels are 1. The uncertain points are then split into three groups: $G_1$ contains points which have a low $\pi$ value[†] with both models, $G_2$ contains points which have a high $\pi$ value[‡] with both models, and all other points are in $G_3$[§]. The algorithm assumes that the most extreme models can be found by giving all the points in $G_1$ the same label, all the points in $G_2$ the same label and only finding all possible combinations of labels for the points within $G_3$. This algorithm reduces the number of logistic regression models fitted to $2 + 2^{2+q'}$ where $q'$ is the number of points in $G_3$.

### 5.3.1 Alternative Methods

#### Exclude Uncertain Results

The most straightforward approach to dealing with this is to remove the uncertain results from $D$ to produce $D_\times$ and a precise logistic regression model $\mathcal{LR}_\times(D)$. This approach may

[†] ALWAYS($\pi \leq 0.5$).

[‡] ALWAYS($\pi > 0.5$).

[§] $0.5 \in \pi$.

**Algorithm L1:** Algorithm to find $\mathcal{ILR}\left(D\right)$ if $D$ has interval uncertainty within its labels.

```
Input: $d = \{(\mathbf{x}_i, y_i)\ \forall i = 0, \ldots, p\}$; $u = \{(\mathbf{x}_j, [0,1])\ \forall j = 0, \ldots, q\}$; $D = d \cup u$
$\mathcal{ILR}\left(D\right) \leftarrow \{\}$;
for all combinations $C \in \{(0, \ldots, 0), (0, \ldots, 1), (1, \ldots, 0), (1, \ldots, 1), \cdots\}$ do
    $D' \leftarrow d \cup \{(\mathbf{x}_j, C_j)\ \forall \mathbf{x}_j \in u\}$;
    $\mathcal{ILR}\left(D\right) \leftarrow \mathcal{ILR}\left(D\right) \cup \mathcal{LR}\left(D'\right)$;
end
Output: $\mathcal{ILR}\left(D\right)$
```

**Algorithm L2:** Algorithm to find $\mathcal{ILR}\left(D\right)$ if $D$ has interval uncertainty within its labels using heuristics to reduce the number of iterations needed.

```
Input: $d = \{(\mathbf{x}_i, y_i)\ \forall i = 0, \ldots, p\}$; $u = \{(\mathbf{x}_j, y_j = [0,1])\ \forall j = 0, \ldots, q\}$; $D = d \cup u$
$D_{(1,\ldots,1)} \leftarrow d \cup \{(\mathbf{x}_j, 1)\ \forall \mathbf{x}_j \in u\}$;
$D_{(0,\ldots,0)} \leftarrow d \cup \{(\mathbf{x}_j, 0)\ \forall \mathbf{x}_j \in u\}$;
Find $\mathcal{LR}\left(D_{(1,\ldots,1)}\right)$ and $\mathcal{LR}\left(D_{(0,\ldots,0)}\right)$;
$G_1 \leftarrow \{\}$; $G_2 \leftarrow \{\}$; $G_3 \leftarrow \{\}$;
for all $u_j = (\mathbf{x}_j, y_j) \in u$ do
    if $\pi_{\mathcal{LR}\left(D_{1,\ldots,1}\right)}(\mathbf{x}_i) < 0.5$ and $\pi_{\mathcal{LR}\left(D_{0,\ldots,0}\right)}(\mathbf{x}_i) < 0.5$ then
        $G_1 \leftarrow G_1 \cup \{u_j\}$
    else if $\pi_{\mathcal{LR}\left(D_{1,\ldots,1}\right)}(\mathbf{x}_i) > 0.5$ and $\pi_{\mathcal{LR}\left(D_{0,\ldots,0}\right)}(\mathbf{x}_i) > 0.5$ then
        $G_2 \leftarrow G_2 \cup \{u_j\}$
    else
        $G_3 \leftarrow G_3 \cup \{u_j\}$
    end
end
for all $A$ in $\{0,1\}$ do
    for all $B$ in $\{0,1\}$ do
        for all combinations $C \in \{(0, \ldots, 0), (0, \ldots, 1), (1, \ldots, 0), (1, \ldots, 1), \cdots\}$ do
            $D' \leftarrow d \cup \{(\mathbf{x}_j, A)\ \forall (\mathbf{x}_j, y_j) \in G_1\} \cup \{(\mathbf{x}_j, B)\ \forall (\mathbf{x}_j, y_j) \in G_2\} \cup$
              $\{(\mathbf{x}_j, C_j)\ \forall (\mathbf{x}_j, y_j) \in G_3\}$;
            $\mathcal{ILR}\left(D\right) \leftarrow \mathcal{ILR}\left(D\right) \cup \mathcal{LR}\left(D'\right)$;
        end
    end
end
Output: $\mathcal{ILR}\left(D\right)$
```

be valid if the missing data is small compared to the total dataset size and if it is missing (completely) at random.

#### Semi-Supervised Logistic Regression

Semi-supervised learning methods extend supervised learning techniques to cope with additional unlabelled data. Numerous authors present semi-supervised logistic regression methods based on a variety of different methods: Amini and Gallinari [254] use Classification Expectation Maximisation; Krishnapuram et al. [263] and Chi et al. [256] use Bayesian methods; Bzdok et al. [264] use an autoencoder and a factored logistic regression model; Chai et al. [265] combine active learning and semi-supervised learning to " achieve better performances compared to the widely used semi-supervised learning and active learning methods." In all cases, it is important that the smoothness, clustering and manifold assumptions are valid to use semi-supervised learning techniques [255].

For this analysis, we have used the *scikit-learn*'s semi-supervised learning algorithm, which uses Yarowsky's algorithm to enable the logistic regression to learn from the unlabelled data [232, 266].

### 5.3.2 Demonstration

Dataset $D_3$ has been created from dataset $D_1$ by replacing five labels from the dataset with the $[0, 1]$ interval. The labels that have been changed are around the point at which the data goes from 0 to 1. This dataset is shown in Figure 5.20, with the uncertain labels plotted as vertical lines. The figure shows all the logistic regression models that have been fitted using both Algorithms L1 (grey lines) and the boounds produced by Algorithm L2 (blue lines). Since the blue lines always correspond to the extremum of the grey lines, Algorithm L2 has correctly estimated the imprecise bounds, and any interval $\pi$ value found from imprecise models is guaranteed to contain the true value.

Figure 5.21 shows the imprecise logistic regression model that is trained on this uncertain dataset, and for comparison, the model trained on the dataset with the dunno labels removed, $\mathcal{LR}_\times(D_3)$, and the semi-supervised model, $\mathcal{LR}_{ss}(D_3)$. From the figure, it is notable that $\mathcal{LR}_\times(D_3)$ and $\mathcal{LR}_{ss}(D_3)$ are similar.

As in Section 5.2.3, it is helpful to consider different scenarios within which the labels have been removed. Figure 5.22 shows four different scenarios within which the data has been made uncertain, and the imprecise logistic regression models have been fitted using Algorithm L2. Using this algorithm allowed plot (b) to be computed since Algorithm L1 would have required $2^{20}$ models to be fitted and have been computationally prohibitive.

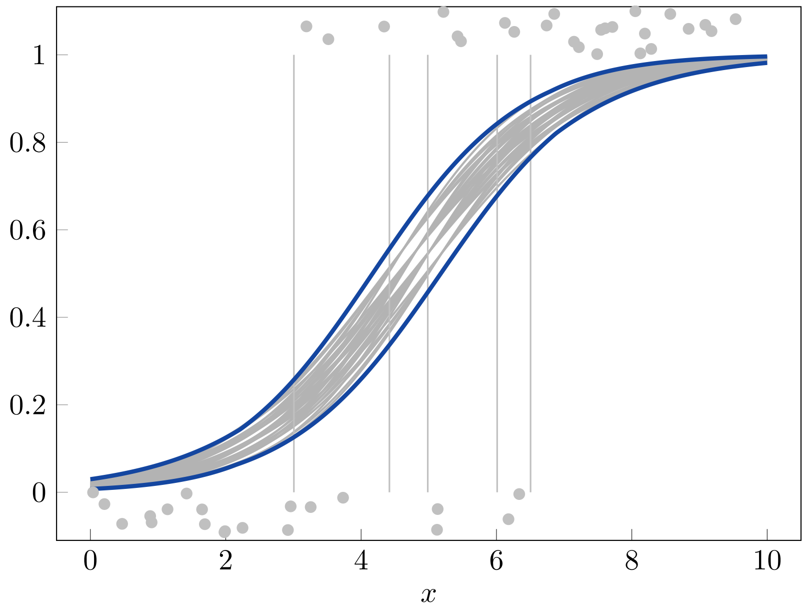


Figure 5.20: Imprecise logistic regression model with the uncertain labels represented by the vertical lines. The grey lines represent all possible models found using Algorithm L1. The blue lines are the bounds found using Algorithm L2.

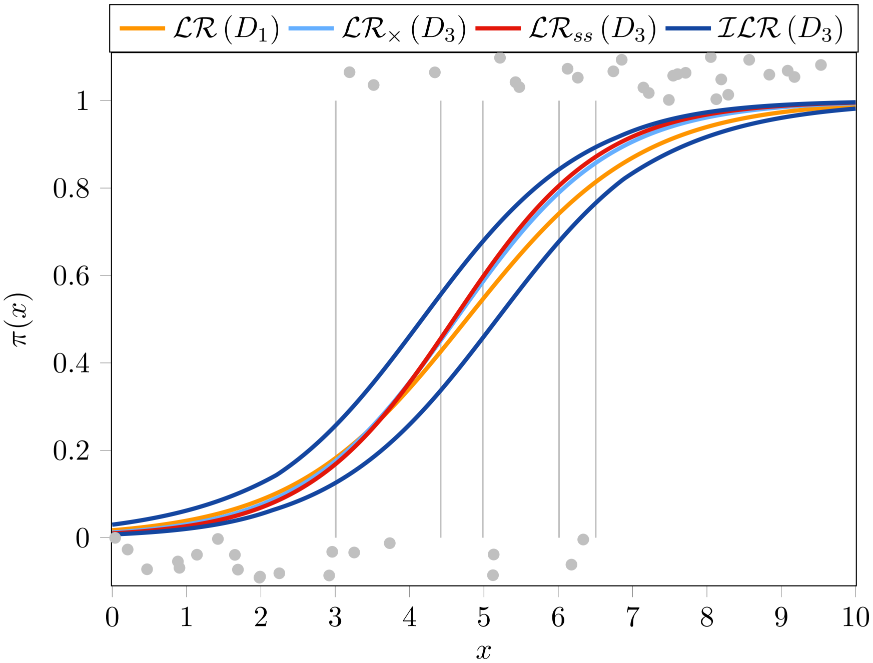


Figure 5.21: Bounds for the imprecise logistic regression or all 50 points with 5 points made uncertain. Compared with $\mathcal{LR}_\times(D_3)$ and $\mathcal{LR}_{ss}(D_3)$.

In all four plots, the imprecise model bounds the base model. It is also notable that the semi-supervised and discarded data approaches are similar in all the plots.

This plot demonstrates that if it can be assumed that the labels are missing (completely) at random, as in (a) and (b), the two alternative approaches are reasonably close to the true model. However, if this is not the case, then there are significant differences between the approaches and the imprecise method must be used to obtain a model that is guaranteed to bound the true model.

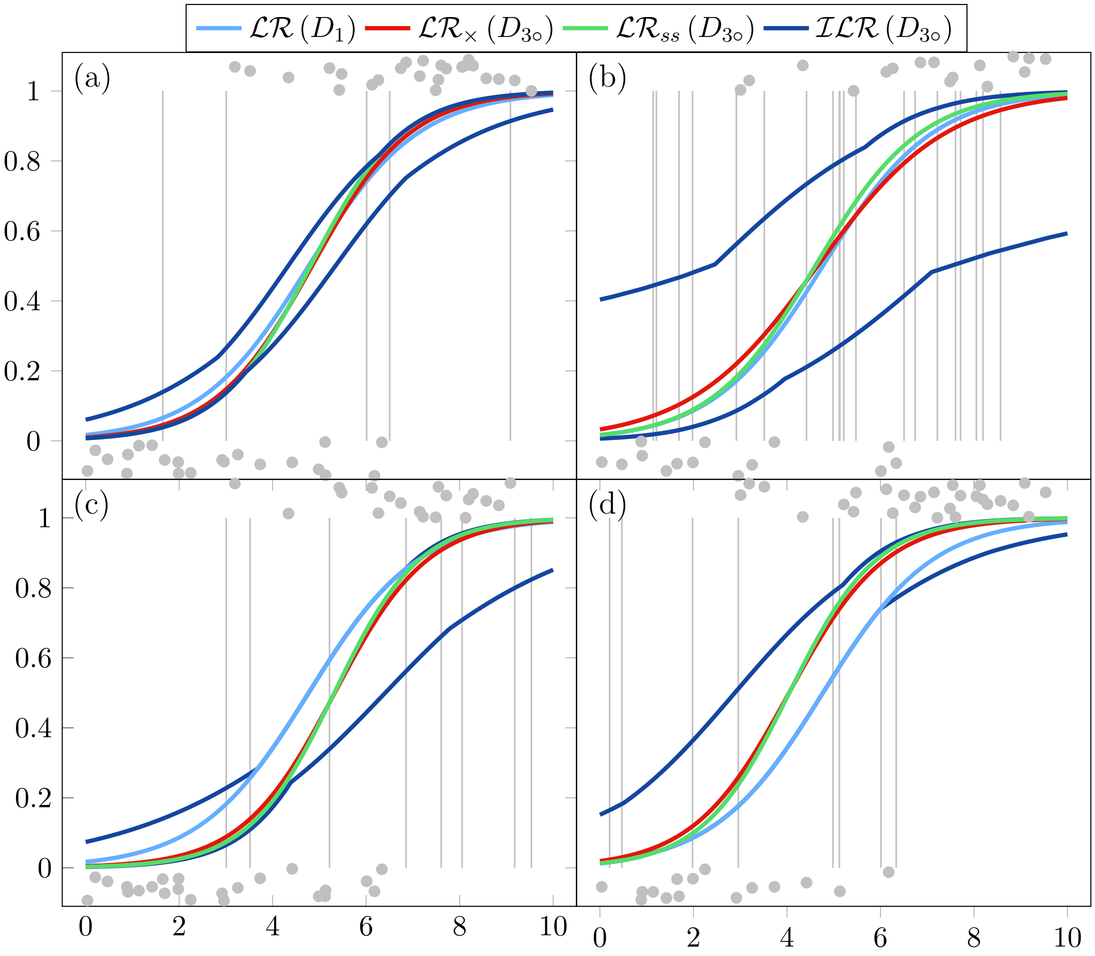


Figure 5.22: Four different scenarios within which dataset $D_1$ has had labels removed. $D_{3a}$ has five labels missing at random. $D_{3b}$ has ten labels missing at random. $D_{3c}$ has eight 1 labels missing. $D_{3d}$ has eight 0 labels missing.

### 5.3.3 White Wine Example

As in Section 5.2.4 it is useful to consider this methodology on a real dataset. In this instance, we will use the white wine dataset from Cortez et al. [262, 267]. This dataset contains the same covariates as the red wine dataset used in Section 5.2.4 but contains many more samples (4898). Again 'good' wine has been defined as having a quality $\geq 6$. This data has been split into training and test samples, with 1618 and 3281 samples, respectively. In order to simulate sommeliers being unsure about the classification of marginally good wine, 100 samples with quality $= 6$ have had their labels removed. Let $W$ be the uncertain dataset and $W^\dagger$ is the original dataset. Algorithm L2 can be used to fit the imprecise logistic regression model on this dataset. For comparison, $\mathcal{LR}_{ss}\left(W\right)$ and $\mathcal{LR}_{\times}\left(W\right)$ have also been found.

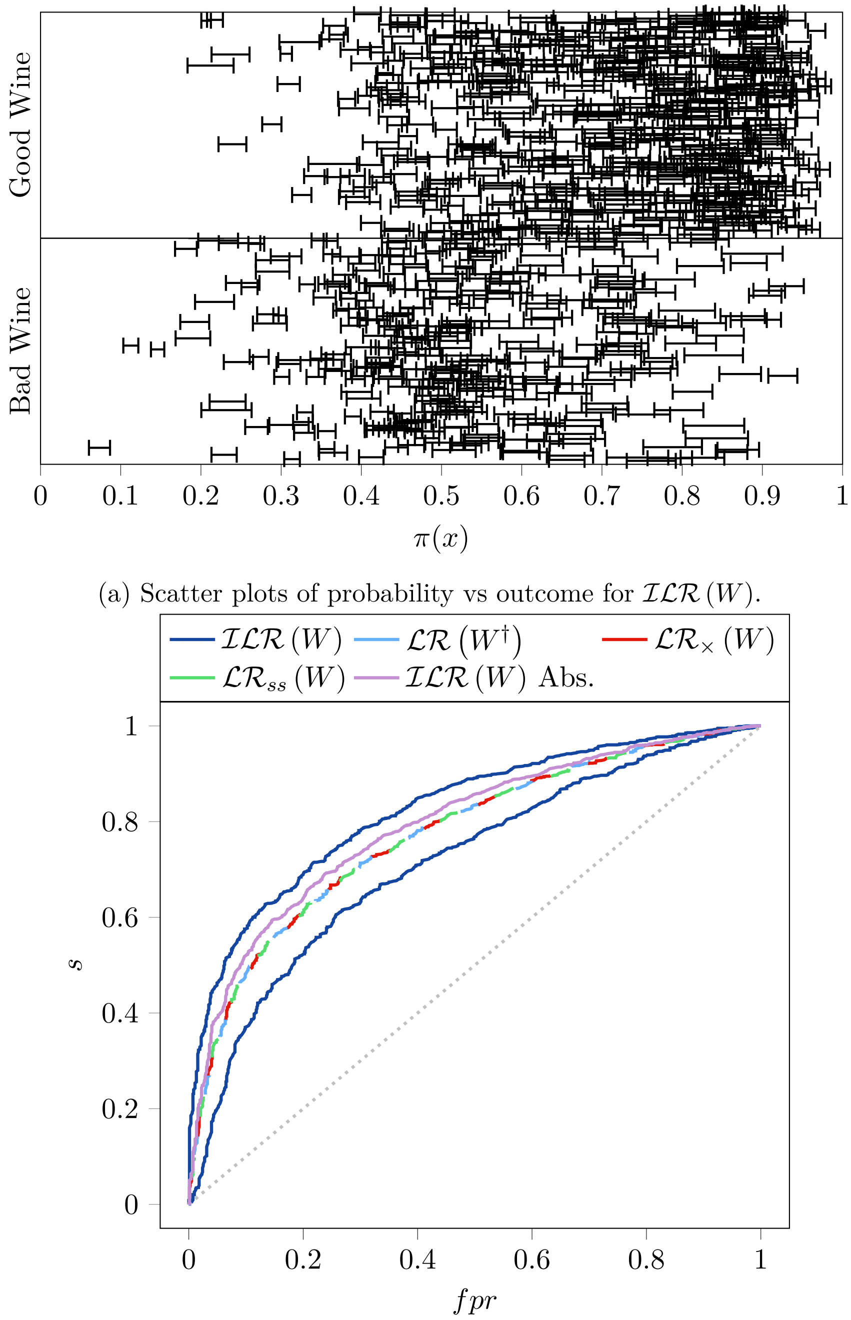
The discrimination plots for these models are shown in Figure 5.23. Figure 5.23a shows that very few points have been given a low probability of being bad wine. Most of the bad wine has $\pi \approx 0.5$ whereas good wine has a high probability ($\pi \approx 0.9$). This plot suggests that when making classifications from the model, selecting a threshold value of $C = 0.7$ would be an appropriate choice to distinguish between the two classes. ROC curves can also be plotted. As with the previous examples, the precise models all have very similar curves and AUC values which the abstaining model 'beats'. In this case:

$$AUC\left[\mathcal{ILR}\left(W\right)\right] = [0.716, 0.826],$$

$$AUC\left[\mathcal{ILR}\left(W\right)(\mathit{Abstain})\right] = 0.794,$$

and

$$AUC[\mathcal{LR}_{ss}\left(W\right)] \approx AUC[\mathcal{LR}_{\times}\left(W\right)] \approx AUC[\mathcal{LR}\left(W^\dagger\right)] = 0.774.$$



(a) Scatter plots of probability vs outcome for $\mathcal{ILR}\left(W\right)$.

(b) Receiver operating characteristic curve for the simple example with added uncertain classifications.

Figure 5.23: Plots to show the discriminatory performance of the logistic regression models for the white wine example.

## 5.4 Uncertainty in Both Features and Labels

The imprecise approach can be used when there is uncertainty about both the features and the labels. Such situations are present in numerous real-world datasets. An imprecise logistic regression model can be found in this scenario through a combination of the algorithms in Sections 5.2 and 5.3.

### 5.4.1 Example

Osler et al. [61] use a logistic regression model to predict the probability of death for a patient after a burn injury. The model they use is based upon a subset of data from the American Burn Association's National Burn Database[¶]. The dataset has a mix of discrete (gender, race, flame involved in injury, inhalation injury) and continuous variables (age, percentage burn surface area) that can be used to model the probability that a person dies (outcome 1) after suffering a burn injury. Osler et al. excluded some patients from the dataset before training their model. They removed patients if their age or 'presence of inhalation injury' was not recorded. Additionally, as patients older than 89 were assigned to a single age category in the original dataset, they gave them a random age between 90 and 100 years.

Osler et al. did not need to exclude these patients merely because of epistemic uncertainty about the values. The methods proposed within this chapter can be used with the original imprecise data. For instance, patients for which the outcome was unknown could have been included within their analysis as described in Section 5.3. Similarly, patients whose inhalation injury or age was unknown could have been included with the method described in Section 5.2. Patients with unknown inhalation injury could have been included as the $[0, 1]$ interval. Patients whose age was completely unknown could have been replaced by an interval between the minimum and maximum age, whereas if there was uncertainty because they were over 90 years old, then they could be intervalised as $[90, 100]$.

Other interval uncertainties may be present within the dataset. It is unlikely to be the case that all the people used within the study fit neatly into the discrete variables given, for instance the variable race is valued at 0 for "non-whites" and 1 for "whites". However, it goes without saying that the diversity of humanity does not fall into such overly simplified categories; there are likely to be many people who could not be given a value of 0 or 1 and should instead have a $[0, 1]$ value. The same is true for gender. Not everyone can be defined as male or female. Also, there is almost certainly some measurement uncertainty associated with calculating the burn surface area that may also be best expressed as intervals. For

[¶] http://ameriburn.org/research/burn-dataset/.

simplicity, these uncertainties have not been addressed below.

For this analysis, the subsample of the dataset used by Osler et al. made available by Hosmer Jr et al. [62, p. 27] has been used. This version of the dataset includes 1000 patients from the 40,000 within the entire study and has a much higher prevalence of death than the original dataset. Because access to the original data is prohibitively expensive, the values in this dataset have been re-intervalised to replicate some of the removed uncertainty to create a hypothetical dataset, $B$, for this exposition. As there are no individuals older than 90 within the dataset, that particular re-intervalisation has not been possible, so all patients older than 80 have had their ages intervalised as [80,90]. Similarly, for 20 patients, the censored inhalation injury has been restored to dunno interval. Ten patients, who had been dropped because their outcome status was unknown, have been restored with status represented as [0,1].

There are two possible routes for an analyst to proceed when faced with such a dataset. They could follow the original methodology of Osler et al. and randomly assign patients with interval ages a precise value and then discard all other patients for which there is some uncertainty. Alternatively, the analyst could include the uncertainty within the model by creating an imprecise logistic regression model. As there is uncertainty within both the features and the labels, the model can be estimated by finding the values within the intervals that correspond to the minimum and maximum for $\beta_0$, $\beta_1$, etc. $\mathcal{ILR}(B)$ is the imprecise logistic regression fitted from this burn data. For comparison, $\mathcal{LR}(B_\times)$ has also been fitted based on removing the uncertainty in $B$ using the same methodology as Osler.

Regarding the performance of the two models, we can again turn to visualisations, as shown in Figures 5.24 and 5.25. Firstly, looking at Figure 5.25 we can see that the vast majority of patients who were given a low probability of death ($\pi$) did indeed survive, and patients who were given a high probability of death died. The ROC plots are shown in Figure 5.24; Figure 5.24a shows the upper right corner of the plot in more detail. $\mathcal{ILR}(B)$ has $AUC = [0.955, 0.974]$, the no prediction model has $AUC = 0.972$ and $\mathcal{LR}(B)$ has $AUC = 0.966$.

It is pertinent to consider how a model is likely to be used and how uncertainty about the predicted probability of death impacts the classification. One method of dealing with this uncertainty that arises in Sections 5.2 and 5.3 is not making a prediction when the interval for $\pi$ straddles $C$. This method may not be appropriate in this example. What should happen when the model is unable to make a prediction should depend on what the result of deciding a patient has a high-risk of death means clinically. If the model was being used to triage patients that need to go to a major trauma centre because the probability of death is

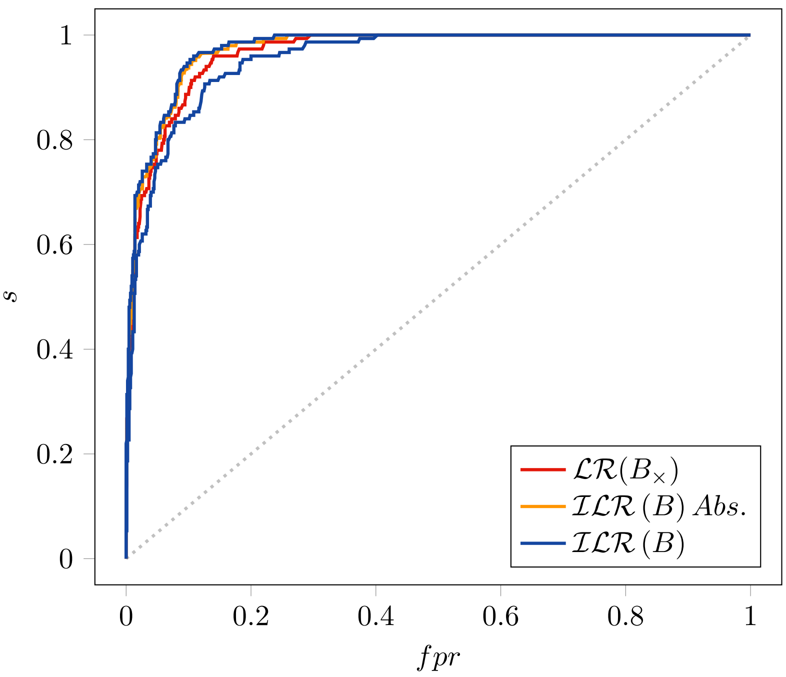


Figure 5.24: Receiver operating characteristic curves for the burn dataset example.

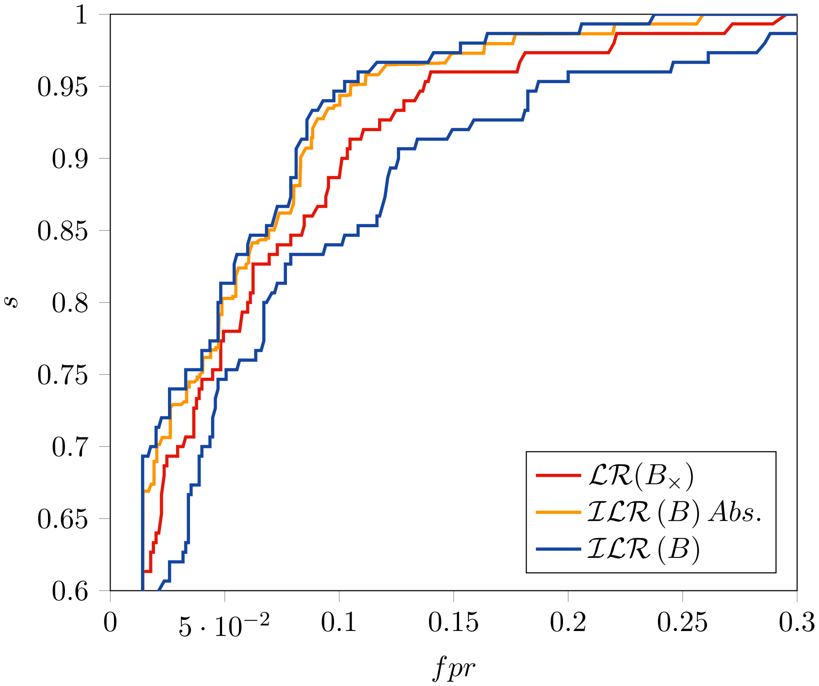


Figure 5.24a: Zoomed version of 5.24

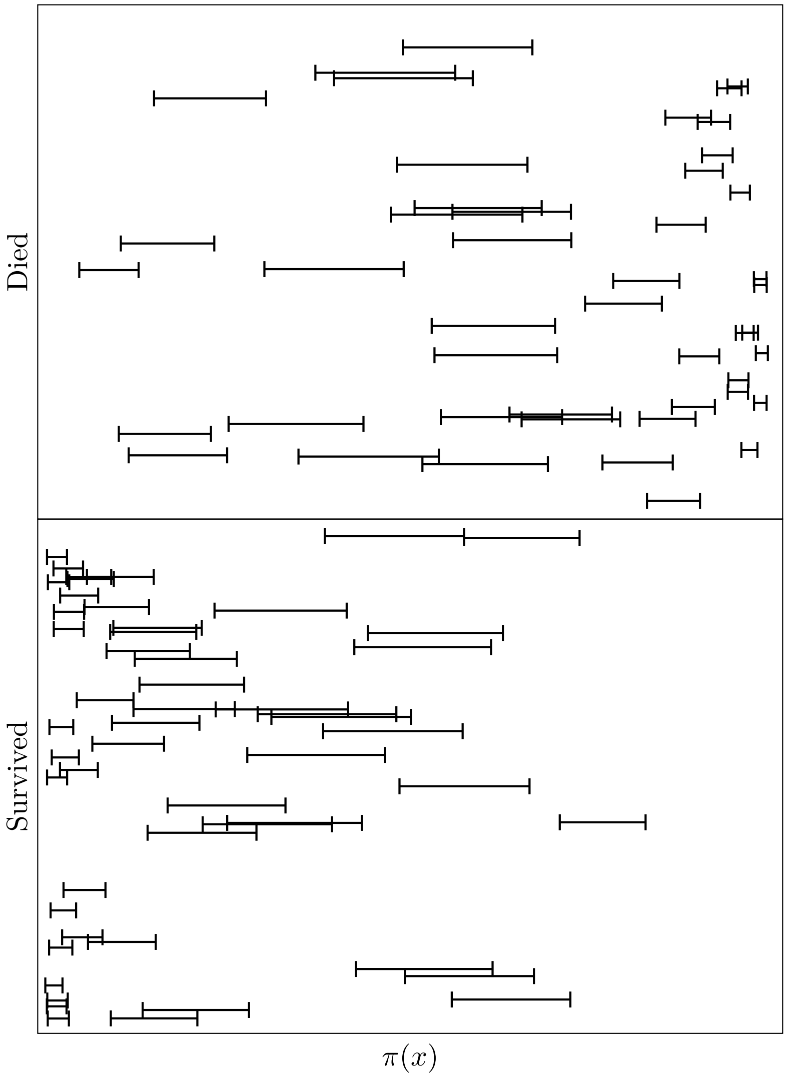


Figure 5.25: Scatter plots of probability vs outcome. The two outcomes have been separated into different plots for clarity.

considered high, then – out of an abundance of caution – one might prefer that if any part of the interval probability was greater than some threshold, the patient should be considered high-risk. This is equivalent to taking the probabilities from the upper bound of the range,

$$\text{high-risk} = \begin{cases} 1, & \text{if } \overline{\pi} \geq C \\ 0, & \text{otherwise.} \end{cases} \tag{5.27}$$

However, if patients who are considered high-risk then undergo some life-altering treatment that is perhaps only preferable to death, then under the foundational medical aphorism of "first do no harm", it may be preferable to consider a patient high-risk only if the whole interval is greater than the decision threshold, this is equivalent of taking the probabilities from the lower bound of the range,

$$\text{high-risk} = \begin{cases} 1, & \text{if } \underline{\pi} \geq C \\ 0, & \text{otherwise.} \end{cases} \tag{5.28}$$

Using the imprecise model in these scenarios would lead to better outcomes as the epistemic uncertainty would not be ignored. It is also the case that if a patient that has a wide interval (as is the case for some in Figure 5.25), implying that there is large epistemic uncertainty about the prediction, the medics would be aware of the uncertainty associated with the model and therefore may prefer to decide another way.

## 5.5 Discussion

Analysts face uncertainties in the measurement values of their data. Often this uncertainty is either assumed away or just completely ignored. However, it may be better to compute with what we know rather than to make assumptions that may need to be revised later. Many uncertainties are naturally expressed as intervals arising from measurement errors and missing or censored values. In the case of logistic regression, when faced with interval uncertainties, samples are often dropped from analyses – assuming that they are missing (completely) at random – or arbitarily replaced by a single value. Analysts should not simply throw this uncertainty away to make subsequent calculations easier.

Interval uncertainties can be included within logistic regression models by considering the set of possible models as an imprecise structure, including in situations where there is uncertainty about the binary outcome status. The present analysis showed that it is not reasonable to throw away data when the status is unknown if the reason the data has gone missing is dependent on the value or status of the missing samples. The case studies showed that previous methods used to handle interval uncertainties are ill-suited to situations where the narrow assumptions that they rely upon are untenable or unjustified. The methods based upon imprecise probabilities described in this chapter work whenever there are interval uncertainties in the data, regardless of how they happened to arise.

The imprecise approach presented within this chapter introduces two distinct uncertainties within logistic regression models. The first is the uncertainty about what the expected label for a particular $\mathbf{x}$ should be, expressed in an aleatoric way by considering $\pi(\mathbf{x})$ as a probability. The second type of uncertainty, added by imprecise logistic regression models, is the epistemic uncertainty expressed by the interval $\pi(\mathbf{x})$ values.

Some may think that, in the case of label uncertainty, it seems counterintuitive that adding unlabelled examples yields an important uncertainty, despite the fact that the unlabelled examples appear uninformative and thus useless. However, it is essential to consider that there is a critical distinction between not knowing a label for some arbitrary value and knowing that we do not know a label for a directly observed value. In the latter case, there is information in the fact that *it is known* that the label is unknown. As shown by Figure 5.22, if an analyst can not assume that the labels are missing (completely) at random, they cannot say that they are uninformative.

Additionally, it is not the case that these uncertain points are being 'added'. They are not being removed or assumed away, as would have been the case in traditional analyses. Previously, analysts did not have methods to characterise this epistemic uncertainty, whereas the presented methods enable analysts to keep their interval data and not neglect them. The situation may occur when one dataset contains precisely known values and another dataset with interval data. Whilst it may seem obvious to pool this data into a big dataset, it may not be the case that bigger data is better data, especially if there is more uncertainty in the pooled dataset. For a full discussion about combining datasets with different levels of imprecision, see Tretiak and Ferson [268].

If the uncertainty within the dataset can be assumed to be missing (completely) at random, the intervals are small compared to the underlying size of the data, or the equidistribution hypothesis holds, analysts may be justified in not performing an imprecise logistic regression and using one of the alternative methods reviewed in this chapter. However, if there is doubt about whether these assumptions hold, then the methods like those presented within here should be used to characterise this uncertainty. It has always been easy to get wrong answers; thus, whilst the algorithms presented within this chapter are computationally expensive, only they will account for the full uncertainty within the dataset.

When using the new approach to classify, each new sample gets an interval probability of belonging to one of the binary classifications. Therefore, there are likely to be samples for which a definitive prediction cannot be made. If an analyst is happy to accept a *don't know* result, then the regression's performance as a classifier may be improved for the samples for which a prediction is made. It may seem counterproductive or unhelpful for a model to return a don't know result. However, this can be desirable behaviour; saying "I don't know" is perfectly valid in situations where the uncertainty is large enough that a different decision could have been reached. As dicussed in Chapter 1, uncertainty in the output can allow for decisions made by algorithms to be more humane by requiring further interrogation to make a classification. Alternatively, depending on the use case, other ways of making decisions based on uncertain predictions could be made, such as being conservative or cost-minimising. Although deciding indeterminate predictions at random would be capricious.

Within many active learning systems, the model is already forced to abstain from predicting labels for samples when the probability is close to the decision boundary ($\pi(x) \approx C$) so that a human can provide a classification [265, 269]. If the imprecise logistic regression model presented in this chapter were used in such a system, it would have the advantage of clearly providing a criteria for which abstentions are prefered, as opposed to a post-hoc decision based upon an arbitrary definition of how close to the boundary is too close. Additionally, this method exposes samples for which the model returns broad interval probabilities even if their centre is not close to the decision boundary and would have been considered a clear decision previously. If an abstention region is provided, then any interval probability produced from the model that straddles the region would be indeterminate.

This work has shown that it is possible to include interval uncertainty in both outcome status and predictor variables within logistic regression analysis by considering the set of possible models as an imprecise structure. Such a method clearly can express the epistemic uncertainties within the dataset that are removed by traditional methods. Future work should be invested in finding improved algorithms to make them less computationally expensive for large-scale datasets.

# Chapter 6

# Impact of Testing Uncertainty in COVID

***Preface*** This chapter is a modified version of N. Gray, D. Calleja, A. Wimbush, E. Miralles-Dolz, A. Gray, M. De Angelis, E. Derrer-Merk, B.U. Oparaji, V. Stepanov, L. Clearkin, and S. Ferson. Is “no test better than a bad test”? Impact of diagnostic uncertainty in mass testing on the spread of COVID-19. *PLoS ONE*, 15(10), 2020. doi: 10.1371/journal.pone.0240775.

The model used in this chapter is a new agent-based SIR model instead of the binomial SIR model used within the original paper. The new model is better able to characterise the dynamics of the mass testing strategies employed in the United Kingdom.

At the start of the COVID-19 pandemic in the United Kingdom, the government’s epidemic management strategy was influenced by epidemiological modelling conducted by several research groups [270, 271]. The analysis of possible mitigation and suppression strategies influenced the government’s approach throughout the pandemic. In March 2020, the lockdown of large parts of society, including school closures, quarantining of infected individuals (self-isolation) and social distancing of the entire population, was considered the “only viable strategy” to suppress the pandemic [272, 273]. Similar analyses in other countries led to over half the worlds population in some form of lockdown by April 2020, and over 90% of global schools closed [274, 275].

When these measures were introduced, it was apparent that the eventual relaxation of lockdown measures was going to be problematic [272]. It was also evident at the time of the first lockdown that the risk of a second wave was significant without a considered cessation of the suppression strategies. Moreover, this wave would possibly be of greater magnitude than the first as the SARS-CoV-2 virus would be endemic in the population [276, 277]. This eventually turned out to be accurate, with further full lockdowns in November 2020 and January 2021, with the end of all restrictions finally occurring in February 2022 [278].

Throughout the pandemic, testing was a key pillar employed by governments to help guide public policy. During the pandemic's initial stages, much attention was focused on the number of tests being conducted [279, 280]. With the direction of the World Health Organisation declaring the need to “test, test, test” [281]. Governments strived to increase the number of tests they could perform rapidly. For instance, on 2nd April 2020, the then Secretary of State for Health, now reality TV ‘star’, Matt Hancock, announced that the UK would increase the number of daily tests to 100,000 per day by the end of that month and that “No test is better than a bad test” [282, 283]. In this chapter, the validity of this claim is explored.

As discussed in Chapter 4, all tests are imperfect, including those for COVID-19. With the hysteria about countries not performing enough tests early in the pandemic, not enough attention was given to the issues of imperfect testing, despite evidence suggesting they are epidemiologically significant [284–287]. The failure to detect the virus in infected patients can be a significant problem in high-throughput settings operating under severe pressure [288].

Several tests were used during the pandemic, but they fall into two distinct classes: ‘have you got it?’ tests for detecting active cases, and the ‘have you had it?’ tests for the presence of antibodies, which imply some immunity to COVID-19. Within the UK, two different testing technologies were used for active viral testing: lab-based reverse transcription polymerase chain reaction (PCR) tests and at-home lateral flow tests [289, 290]. Although much hyped when first suggested [291, 292], there was never a widespread antibody testing programme within the UK.

These two classes of tests need to maximise different test characteristics. Active viral tests need to maximise the sensitivity to help control pandemics. High sensitivity reduces the chance of missing people who have the virus, who may then go on to infect others. There is an additional risk that an infected person who has been incorrectly told they do not have the disease, when in fact they do, may behave in a more less conservative manner than if their disease status were uncertain.

The second testing approach, seeking to detect the presence of antibodies to identify those who have had the disease, would be used in a different strategy. This strategy would involve detecting those who have successfully overcome the virus and are likely to have some level of immunity (or at least reduced susceptibility to severe illness if they are infected again), so they are relatively safe to relax their personal social distancing measures. This strategy would require a high test specificity, aiming to minimise how often the test tells someone they have had the disease when they have not [293]. A false positive tells people they have immunity when they do not, which is even worse than when people are uncertain about their viral history. Although not implemented using antibody tests during the COVID-19 pandemic, the immunity passport idea did reappear in discussions about vaccine passports [294, 295], especially for international travel [296].

## 6.1 Evidence COVID-19 Tests Were Imperfect

The early success in 2020 of South Korea, Singapore, Taiwan and Hong Kong in limiting the impact of the SARS-CoV-2 virus was attributed to their ability to deploy widespread testing, with digital surveillance, and impose targeted quarantines in some cases [288]. This testing was predominantly based on the use of PCR tests. During the 2009 H1N1 pandemic, the rapid development of high-sensitivity PCR assayes was employed early with some success in that global pandemic [297]. When well-targeted, these tests provide a valuable tool for managing and tracking pandemics.

Early PCR tests formed the basis of much of the research into the incidence, dynamics and comorbidities of SARS-CoV-2, but few, if any, of these studies considered the impact of false test results [298–302]. It may be that false test results contributed to some of the counter-intuitive disease dynamics observed early in the pandemic [303].

There was evidence that both PCR tests [304–308] and antibody [309–311] tests lacked perfect sensitivity and specificity even in best-case scenarios. Alternative screening methods such as chest x-rays may be found to have high sensitivity based on biased data [312] or may perform poorly even compared to imperfect RT-PCR tests [307]. The Foundation for Innovative New Diagnostics (FIND) evaluated five RT-PCR tests, scoring highly out of 17 candidate tests on criteria such as regulatory status and availability [313]. Even ideal laboratory conditions can produce a specificity which could be as low as 90%, and the real-world specificity is likely to be lower still.

PCR tests for COVID-19 were considered the gold standard active virus test but, as they required laboratory analysis, were expensive and took several days to return a result to an individual suspected of having COVID. As a result, there was a desire to introduce lateral flow tests that could provide results in as little as 15 minutes. These tests were known when first introduced to be imperfect [314]. Whilst there was debate about the efficacy of mass testing [315–317], at various points over 1 million of these tests were performed daily in the UK [318].

## 6.2 SIR Modelling

SIR models offer one approach to explore infection dynamics, and the prevalence of a communicable disease. In the generic SIR model, there are $\mathcal{S}$ people susceptible to the illness, $\mathcal{I}$ people infected, and $\mathcal{R}$ people who are recovered with immunity. Infected people are able to infect susceptible people at rate $\beta$ and they recover from the disease at rate $\gamma$ [319]. Once infected persons have recovered from the disease they are unable to become infected again or infect others. This may be because they now have immunity to the disease or because they have unfortunately died. Within this basic SIR model, the assumption appears to be incorrect for COVID-19 [320].

Deterministically, at time $t$ the number of individuals in each of the states is given by:

$$\mathcal{S}_t = \mathcal{S}_{t-1} - \Delta_{\mathcal{S},\mathcal{I}} \tag{6.1}$$

$$\mathcal{I}_t = \mathcal{I}_{t-1} + \Delta_{\mathcal{S},\mathcal{I}} - \Delta_{\mathcal{I},\mathcal{R}} \tag{6.2}$$

$$\mathcal{R}_t = \mathcal{R}_{t-1} + \Delta_{\mathcal{I},\mathcal{R}} \tag{6.3}$$

where $S_{t-1}$ is the number of suseptable individuals at the previous time set and

$$\Delta_{\mathcal{S},\mathcal{I}} = \frac{\beta \mathcal{S}_{t-1} \mathcal{I}_{t-1}}{\mathcal{S}_{t-1} + \mathcal{I}_{t-1} + \mathcal{R}_{t-1}} \tag{6.4}$$

$$\Delta_{\mathcal{I},\mathcal{R}} = \frac{\gamma \mathcal{I}_{t-1}}{\mathcal{S}_{t-1} + \mathcal{I}_{t-1} + \mathcal{R}_{t-1}}. \tag{6.5}$$

Figure 6.1 shows a schematic of the generic model formulation, and how people move between the states.

The SIR model has two ways in which the number of new infections falls to zero. Either the number of susceptible people reduces to a point at which the disease can no longer propagate, perhaps because of a vaccine or natural immunity, or the epidemic stops if the basic reproduction rate, $R_0$, of the disease falls below 1 due to social distancing or effective viral suppression:

$$R_0 = \frac{\beta}{\gamma}. \tag{6.6}$$

Figure 6.2 demonstrates the typical disease dynamics.

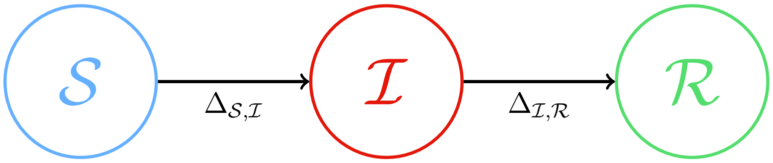

$\mathcal{S}$
$\mathcal{I}$
$\mathcal{R}$
$\Delta_{\mathcal{S},\mathcal{I}}$
$\Delta_{\mathcal{I},\mathcal{R}}$

Figure 6.1: Diagram for a basic SIR model. The black arrows show how people move between the different states.

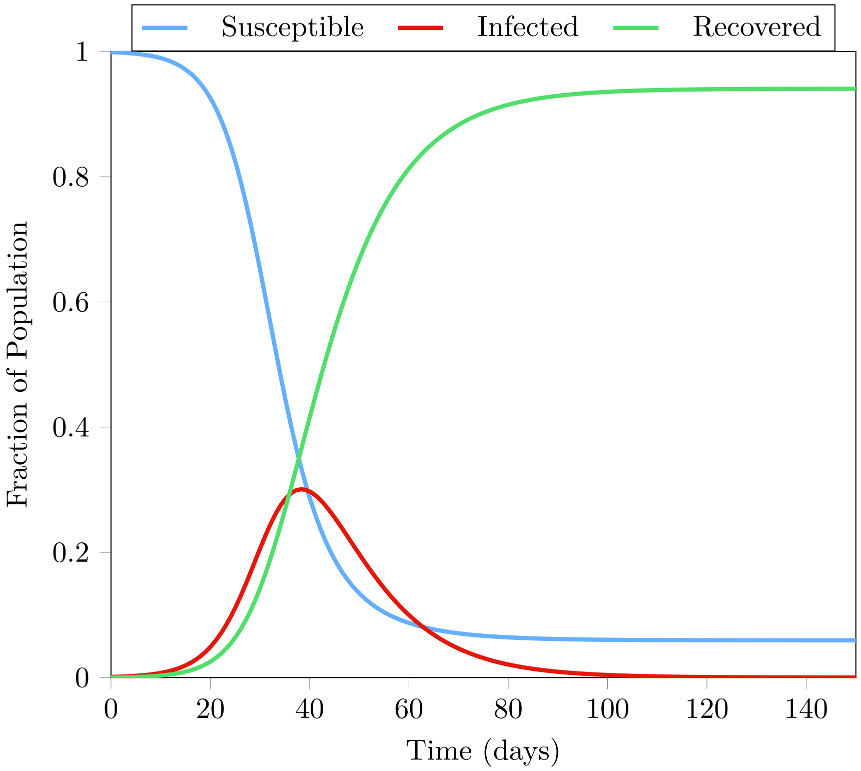


Figure 6.2: Generic SIR model run with $I(t = 0) = 0.0001$, $\beta = 0.3$ and $\gamma = 0.1$.

### 6.2.1 SIR Model with Testing

To explore the effect of imperfect testing on the disease dynamics when strategies are employed to relax the current social distancing measures the classic SIR model has been modified. Three new classes were added to the model: the first is a quarantined susceptible state, $\mathcal{QS}$, the second is a quarantined infected state, $\mathcal{QI}$, and the third is people who have recovered but are in quarantine, $\mathcal{QR}$.

To model lockdown scenarios, the evaluations begin with a majority of the population in the $\mathcal{QS}$ state. Whilst in this state the transmission rate of the disease is totally suppressed. The model evaluates each day's average population-level state transitions. There are two possible tests that can be performed:

- An active virus infection test that is able to determine whether or not someone is currently infectious. This test is performed on some proportion of the unquarantined population $(S + I + R)$. It has a sensitivity of $s_A$ and a specificity of $t_A$.

- An antibody test that determines whether or not someone has had the infection in the past. This is used on the fraction of the population that is currently in quarantine but not infected $(\mathcal{QS} + \mathcal{QR})$ to test whether they have had the disease or not. This test has a sensitivity of $s_B$ and a specificity of $t_B$.

These two tests are used on some of those eligible for testing each day, $\Pi$ infectious people and $\Phi$ susceptible people, up to a total number of tests that can be performed, $C$. A person (in any category) who tests positive in an active virus test transitions into the (corresponding) quarantine state, where they are unable to infect anyone else. A person, in $\mathcal{QS}$ or $\mathcal{QR}$, who tests positive in an antibody test transitions to $S$ and $R$ respectively. There is no transition $\mathcal{R}$ to or from $\mathcal{QR}$ for active virus tests as this would have no impact on the dynamics of the model.

For this parameterisation the impact of being $\mathcal{QS}$ makes an individual insusceptible to being infected. Similarly, individuals in $\mathcal{QI}$ are unable to infect anyone else and remain in that state until they recover from the virus. In reality adherence to full self-isolation was only 42.5% [321], but for the sake of exploring the impact of testing uncertainty these effects are neglected from the model. The participation in infection propagation of individuals in either quarantine state are idiosyncratic, and on average are assumed to be negligibly small for the sake of this analysis.

The deterministic equations for this system are:

$$\begin{aligned}
\mathcal{S}_t &= \mathcal{S}_{t-1} - \Delta_{\mathcal{S},\mathcal{QS}} + \Delta_{\mathcal{QS},\mathcal{S}} - \Delta_{\mathcal{S},\mathcal{I}}, && (6.7)\\
\mathcal{QS}_t &= \mathcal{QS}_{t-1} + \Delta_{\mathcal{S},\mathcal{QS}} - \Delta_{\mathcal{QS},\mathcal{S}}, && (6.8)\\
\mathcal{I}_t &= \mathcal{I}_{t-1} + \Delta_{\mathcal{S},\mathcal{I}} - \Delta_{\mathcal{I},\mathcal{QI}} - \Delta_{\mathcal{I},\mathcal{R}}, && (6.9)\\
\mathcal{QI}_t &= \mathcal{QI}_{t-1} + \Delta_{\mathcal{I},\mathcal{QI}} - \Delta_{\mathcal{QI},\mathcal{R}}, && (6.10)\\
\mathcal{R}_t &= \mathcal{R}_{t-1} + \Delta_{\mathcal{I},\mathcal{R}} + \Delta_{\mathcal{QI},\mathcal{R}} + \Delta_{\mathcal{QR},\mathcal{R}}. && (6.11)
\end{aligned}$$

If the tests were almost perfect, then we can imagine how the epidemic would die out very quickly by either widespread infection or antibody testing with a coherent management strategy. A positive test on the former, and the person is removed from the population; a positive test on the latter and the person, now unable to contract the disease again, can rejoin the population.

More interesting are the effects of incorrect test results on the disease dynamics. If someone falsely tests positive in the antibody test, they enter the susceptible state. Similarly, if an infected person receives a false negative for the disease they remain active in the infected state and hence can infect further people.

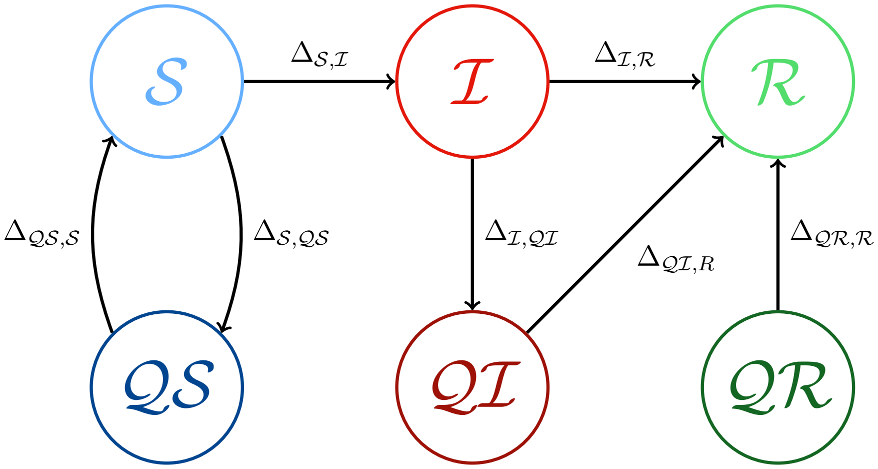

$\mathcal{S}$
$\Delta_{\mathcal{S},\mathcal{I}}$
$\mathcal{I}$
$\Delta_{\mathcal{I},\mathcal{R}}$
$\mathcal{R}$
$\Delta_{\mathcal{QS},\mathcal{S}}$
$\Delta_{\mathcal{S},\mathcal{QS}}$
$\Delta_{\mathcal{I},\mathcal{QI}}$
$\Delta_{\mathcal{QI},\mathcal{R}}$
$\Delta_{\mathcal{QR},\mathcal{R}}$
$\mathcal{QS}$
$\mathcal{QI}$
$\mathcal{QR}$

Figure 6.3: SIR Model used to simulated the effect of mass testing to leave quarantine.

### 6.2.2 Agent Based Model

To more realistically model the effect of testing and quarantining within the simulation, an agent based simulation has been used. Within this $N$ people are individually modelled. Each person is either susceptible, infectious or recovered and may be quarantined. On each day within the simulation there is a probability that each susceptible person becomes infectious,

$$\frac{\beta \mathcal{I}_{t-1}}{\mathcal{S}_{t-1} + \mathcal{I}_{t-1} + \mathcal{R}_{t-1}}. \tag{6.12}$$

The probability that an infectious person recovers is $\gamma$.

On each day, infectious people have a probability $\Pi$ of wanting a test and susceptible people have a probability $\Phi$ of wanting a test. If the number of people that want a test is greater than the testing capacity $C$ then the tests can be targeted. $T$ is the targeting parameter: if $T = 1$ then all tests on infectious people are performed before any on susceptible people, if $T = 0$ then the tests are performed on infectious and suseptable people at random. Algorithm 6.2 specifies how this targeting happens.

If a person is eligible for a test, depending on the test type and result, they may move between the quarantine and active states. If a person receives a positive active virus test then they move to the respective quarantine state, $\mathcal{S} \longrightarrow \mathcal{QS}$ or $\mathcal{R} \longrightarrow \mathcal{QR}$, until their test result expires after $X_+$ days. If they receive a negative test result they stay in their current state and are not eligible to be tested again for $X_-$ days. For antibody tests if a person receives a positive result they leave their quarantine state, $\mathcal{QS} \longrightarrow \mathcal{S}$ etc. Positive antibody tests are valid forever, people with a negative result can retest for antibodies after $X_-$ days.

The full simulation is described in Algorithm 6.1. Figure 6.4 shows a simulated pandemic with $\beta = 0.3$, $\gamma = 0.1$, that starts with $\mathcal{S}_0 = 980$, $\mathcal{I}_0 = 20$ (all other states empty). There is an active virus test with $s = t = 0.9$, $C = 50$, $\Pi = 0.5$, $\Phi = 0.01$ and no targeting. Test results expire after $X_+ = 7$ and $X_- = 3$ days.

## 6.3 Using Testing to Relax Lockdown Restrictions

In order to explore the possible impact of testing strategies on the relaxation of lockdown measures several scenarios have been analysed. Two scenarios are considered:

- **Immediate end to lockdown scenario**: This baseline scenario is characterised by a sudden relaxation of lockdown measures, followed by active virus testing to try to control the spread of the virus.
- **Immunity passports scenario**: Using antibody based testing to identify those who have some level of natural immunity and making those individuals exempt from all personal social distancing rules. This policy was discussed in the media at the start of the pandemic [322–324] but never implemented. The concept is similar to the concept of vaccine passports which were introduced later in the pandemic, especially

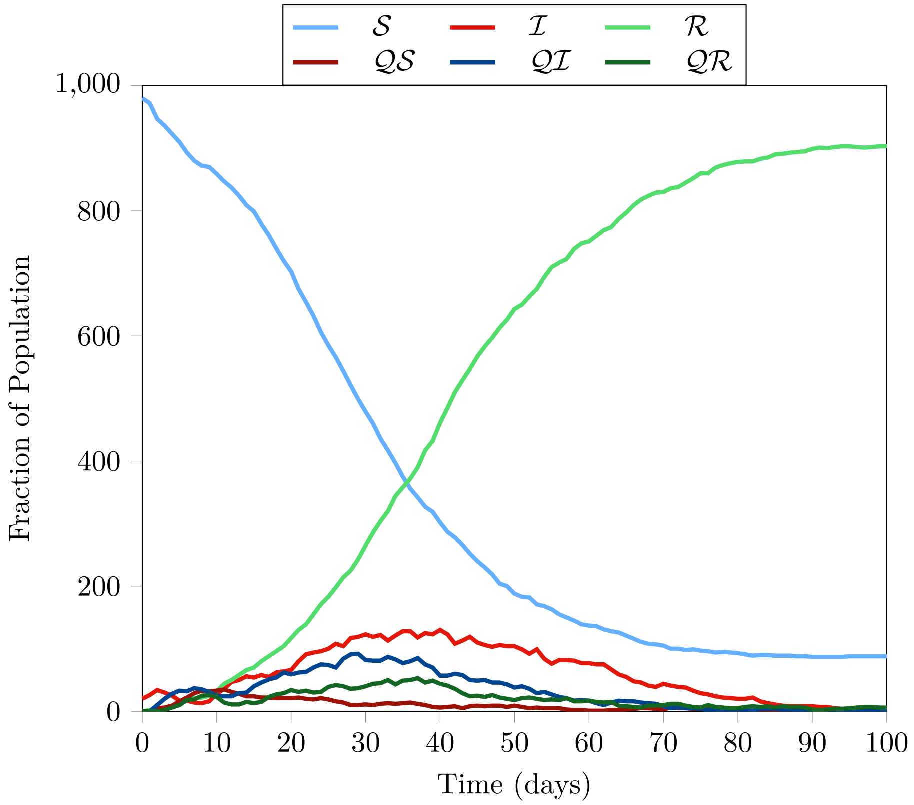


Figure 6.4: A simulated pandemic using the SIRQ model. $\beta = 0.3$, $\gamma = 0.1$, that starts with $\mathcal{S}_0 = 980$, $\mathcal{I}_0 = 20$ (all other states empty) that has a active virus test with $s = t = 0.9$, $C = 50$, $\Pi = 0.5$, $\Phi = 0.01$ and no targeting.

**Algorithm 6.1:** Population update

```
Input: Number of days, $\beta$, $\gamma$
Input: Population: $\mathcal{S}_0$, $\mathcal{I}_0$, $\mathcal{R}_0$, $\mathcal{QS}_0$, $\mathcal{QI}_0$, $\mathcal{QR}_0$
Input: Test Parameters: $C$, $s$, $t$, $\Pi$, $\Psi$, $X_+$, $X_-$, $T$
foreach day do
    $\Delta_{\mathcal{S},\mathcal{I}} = \frac{\beta\mathcal{S}_{t-1}\mathcal{I}_{t-1}}{\mathcal{S}_{t-1}+\mathcal{I}_{t-1}+\mathcal{R}_{t-1}}$;
    for all people $(P) \in$ population do
        if $P \in \mathcal{S}$ then
            if $R \leq \Delta_{\mathcal{I}}$ then $P \longrightarrow \mathcal{I}$;
        if $P \in \mathcal{I}$ or $\mathcal{QI}$ then
            if $R \leq \gamma$ then $P \longrightarrow \mathcal{R}$;
        if $P$ has +ve test result then
            if days since test $> X_+$ then $P$ leaves quarantine;
        if $P$ has -ve test result then
            if days since test $> X_-$ then $P$ leaves quarantine;
    Using Algorithm 6.2 to get population eligible for test;
    Count number of people in $\mathcal{S}_0$, $\mathcal{I}_0$, $\mathcal{R}_0$, $\mathcal{QS}_0$, $\mathcal{QI}_0$, $\mathcal{QR}_0$
```

**Algorithm 6.2:** Selecting which individuals to test

```
Input: Population, $C$, $\Phi$, $\Pi$
for all untested people $(P) \in$ population do
    if $P \in \mathcal{S}$ then
        if $R \leq \Phi$ then
            $P$ wants test;
    if $P \in \mathcal{I}$ then
        if $R \leq \Pi$ then
            $P$ wants test;
Select $T \times \mathcal{I}$ infectious people to go to front of the queue;
First $C$ people are eligible for a test;
Output: Population to be tested
```

**Algorithm 6.3:** Performing Active Virus test with sensitivity $s$ and specificity $t$ on an individual ($P$). $R$ is a random number in $U(0,1)$

```
Input: Test Parameters: $s$, $t$
if $P \in \mathcal{S}$ then
    if $R > t$ then
        $P \leftarrow +ve$ result;
        $P \longrightarrow \mathcal{QS}$
    else
        $P \leftarrow -ve$ result;
if $P \in \mathcal{I}$ then
    if $R \leq s$ then
        $P \leftarrow +ve$ result;
        $P \longrightarrow \mathcal{QI}$
    else
        $P \leftarrow -ve$ result;
if $P \in \mathcal{R}$ then
    if $R > t$ then
        $P \leftarrow +ve$ result;
        $P \longrightarrow \mathcal{QR}$
    else
        $P \leftarrow -ve$ result;
```

**Algorithm 6.4:** Performing Antibody test with sensitivity $s$ and specificity $t$ on an individual ($P$). $R$ is a random number in $U(0,1)$

```
Input: Test Parameters: $s$, $t$
if $P \in \mathcal{QS}$ then
    if $R > t$ then
        $P \leftarrow +ve$ result;
        $P \longrightarrow \mathcal{S}$
    else
        $P \leftarrow -ve$ result;
if $P \in \mathcal{QR}$ then
    if $R \leq s$ then
        $P \leftarrow +ve$ result;
        $P \longrightarrow \mathcal{R}$
    else
        $P \leftarrow -ve$ result;
```

for international travel.

These scenarios are illustrative of the type of impact, and the likely efficacy of a range of different testing configurations. There are significant sociological and ethical considerations associated with any of these approaches and the analysis presented is purely on the question of efficacy.

For the purpose of the analysis we have selected a population similar in size to the United Kingdom, 67 million people. $\beta$ and $\gamma$ were set to 0.32 and 0.1 respectively, to ensure that the $R_0$ value of the model was broadly in line with other models [325, 326].

### 6.3.1 Immediate End to Lockdown Scenario

Under the baseline scenario, characterised by the sudden and complete cessation of lockdown measures, we explored the impact of infection testing. Under this formulation the initial conditions of the model in this scenario are that the all of the people in $\mathcal{QS}$ transition to $\mathcal{S}$ in the first iteration. The impact of infection testing under this scenario was analysed in Figure 6.5 using the parameters shown in Table 6.1.

These scenarios consider the impact of attempts to control the disease through increased testing capacity and a more sensitive test. A test capacity range between 0.001 and 0.05 as a fraction of the population per day were considered. For comparion, the highest number of tests performed on an individual day within the UK was $\approx 0.035$ (2.35 million) on the 5th January 2022 [318]. 0.001 would be representative of the number of tests performed during the initial lockdown in April 2020. To illustrate the sensitivity of the model to testing scenarios an evaluation was conducted with a range of infection test sensitivities, from 0.50 (i.e of no diagnostic value) to 0.98.

As would be expected, the model indicates that an exponential increase in cases is an inevitability and as many as 30% of the population could become infected within 30 days. At higher capacities this pattern remains, though peak infection counts are reduced. This comes with increasing numbers of people being forced to quarantine. Overall it is clear that reliance on infection testing, even with a highly sensitive test and high capacities, is not enough to prevent widespread infection. The initial spike of the number of people in quarantine in the high capacity cases ($C = 0.01$ and $C = 0.05$) are due to people entering quarantine but nobody being released from quarantine since nobody has spent $X_+$ days in quarantine.

The specificity of these tests has a negligible impact on the total number of infections as can be seen in Figure 6.5. A false positive would mean people are unnecessarily removed from the susceptible population, but the benefit of a reduction in susceptible population is

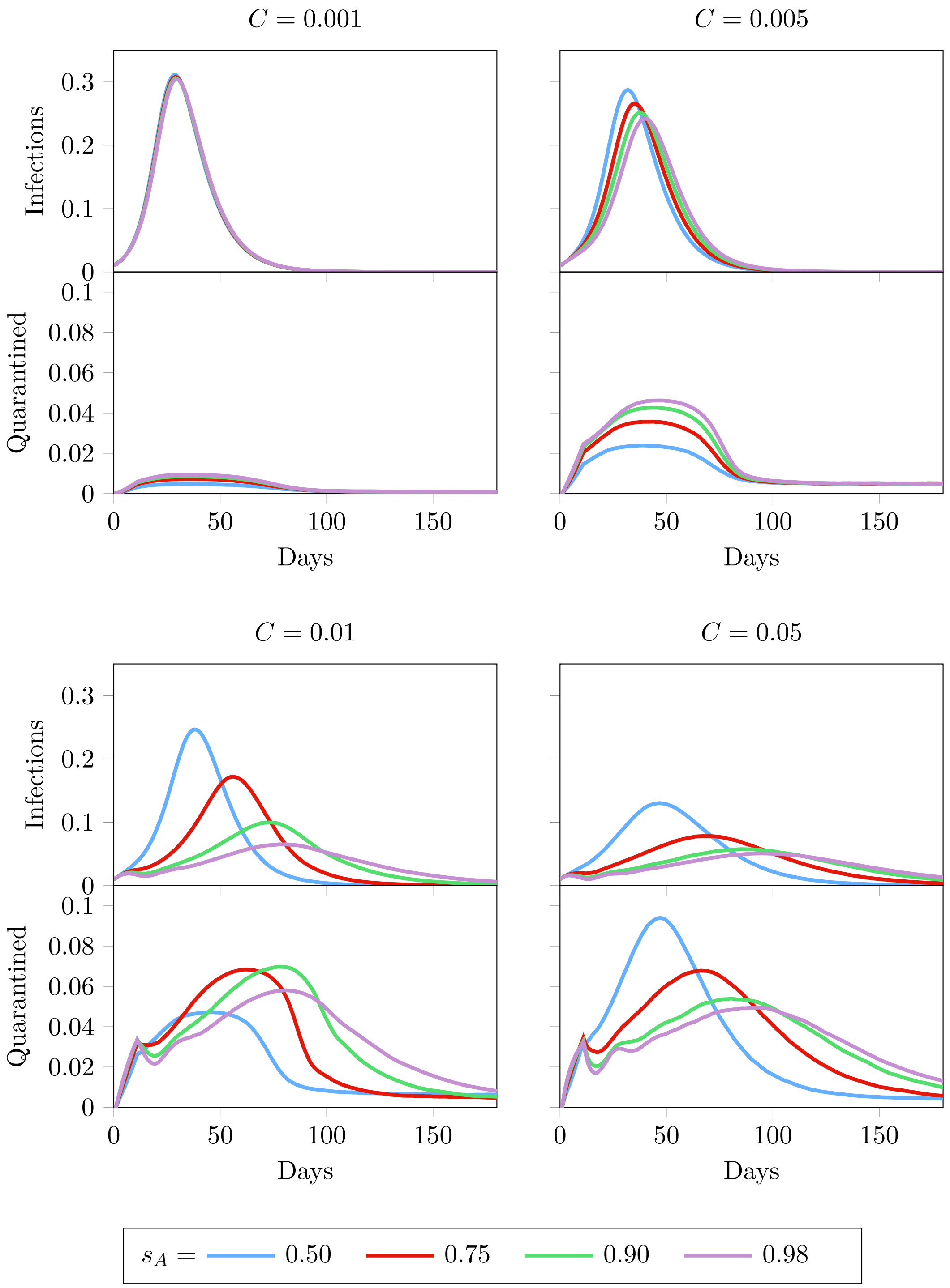


Figure 6.5: A comparison of different test sensitivities $s$ and capabilities $C$. Top: The number of infected individuals $(\mathcal{I} + \mathcal{QI})$. Bottom: The proportion of the population that is quarantined $(\mathcal{QS} + \mathcal{QI} + \mathcal{QR}$ population). Model parameters are shown in Table 6.1.

| Epidemic Parameters | | | | | | | |
|---|---|---|---|---|---|---|---|
| $\beta = 0.32$ | | | | $\gamma = 0.1$ | | | |
| Active Virus Test Parameters | | | | | | | |
| $C$ | $s$ | $t$ | $\Pi$ | $\Psi$ | $X_+$ | $X_-$ | $T$ |
| varies | varies | 0.8 | 0.5 | 0.01 | 10 | 3 | 0 |
| Initial Population split | | | | | | | |
| N | $\mathcal{S}$ | $\mathcal{I}$ | $\mathcal{R}$ | $\mathcal{QS}$ | $\mathcal{QI}$ | $\mathcal{QR}$ | |
| $10^6$ | 0.99 | 0.01 | 0 | 0 | 0 | 0 | |

Table 6.1: Fixed parameters used for Figure 6.5 analysis. Antibody tests were disabled for this analysis.

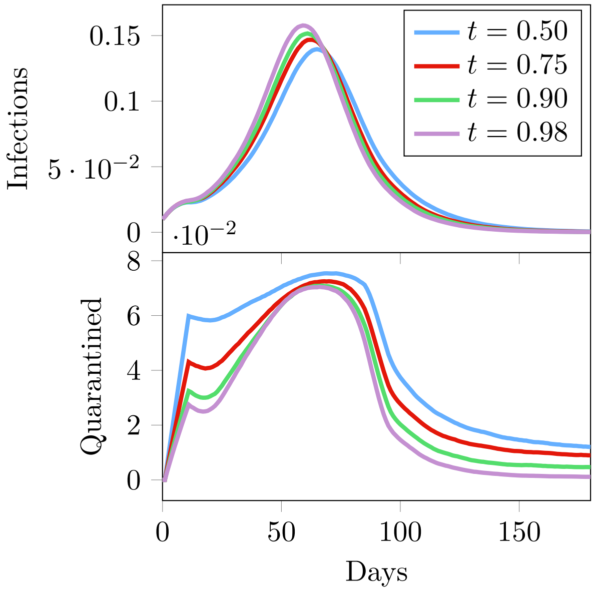


Figure 6.6: A comparison of different test specificities $t$. Model parameters shown in Table 6.2.

| Epidemic Parameters | | | | | | | |
|---|---|---|---|---|---|---|---|
| $\beta = 0.32$ | | | | $\gamma = 0.1$ | | | |
| Active Virus Test Parameters | | | | | | | |
| $C$ | $s$ | $t$ | $\Pi$ | $\Psi$ | $X_+$ | $X_-$ | $T$ |
| varies | 0.8 | varies | 0.5 | 0.01 | 10 | 3 | 0 |
| Initial Population split | | | | | | | |
| N | $\mathcal{S}$ | $\mathcal{I}$ | $\mathcal{R}$ | $\mathcal{QS}$ | $\mathcal{QI}$ | $\mathcal{QR}$ | |
| $10^6$ | 0.99 | 0.01 | 0 | 0 | 0 | 0 | |

Table 6.2: Fixed parameters used for Figure 6.6 analysis. Antibody tests were disabled for this analysis.

negligibly small. In this case a test with a worse specificity reduces the number of infections the most, this is a result of there being more people quarantined.

Figure 6.7 explores the effect of test targeting. This shows that with a small capacity ($C = 0.001$) targeting has a limited impact on the total number of infections. As there are not enough tests to go around and, therefore, even if all tests go to infectious individuals, too many infectious people go undiagnosed and unquarantined. Equally, targeting has a limited impact when the capacity is high ($C = 0.05$), as every person who wants a test can get one (i.e. the capacity is de facto unlimited).

The $C = 0.005$ and $C = 0.01$ plots demonstrate that, if there are more individuals wanting a test than the capacity allows, targeting can have a large impact of the total number of infections. The $C = 0.005/T = 1$, $C = 0.01/T > 0.5$ cases all have similar case numbers to each other, suggesting that there is a threshold at which, if authorities can target tests accurately, increased testing may not be needed to reduce infections and may even be unhelpful as large numbers have to quarantine unnecessarily.

### 6.3.2 Immunity Passport Scenario

The immunity passport is an idiom describing an approach to the relaxation of lockdown measures that focuses heavily on antibody testing. The efficacy of this strategy, however, is far more dependent on the prevalence of antibodies (seroprevalence) in the general population. Presumably, only people who test positive for antibodies would be allowed to leave quarantine. The more people in the population with antibodies, the more that will get a true positive and so would be correctly allowed to leave quarantine (under the paradigms of an immunity passport).

The danger of such an approach are false positives. We demonstrate the impact of people reentering the susceptible population who have no immunity. We assume their propensity to contract the infection is the same as those without the false sense of security a positive test may engender. On an individual basis, and even at the population level, behavioural differences between those with false security from a positive antibody test, versus those who are uncertain about their viral history could be significant. The model parameterisation here does not include this additional confounding effect.

To simulate the seroprevalence in the general population the model is preconditioned with different proportions of the population in the $\mathcal{QS}$ and $\mathcal{QR}$ states. This is analogous to the proportion of people that are currently in quarantine who have either had the virus and developed some immunity, and the proportion of the population who have not contracted the virus and have no immunity. Of course the individuals in these groups do not really

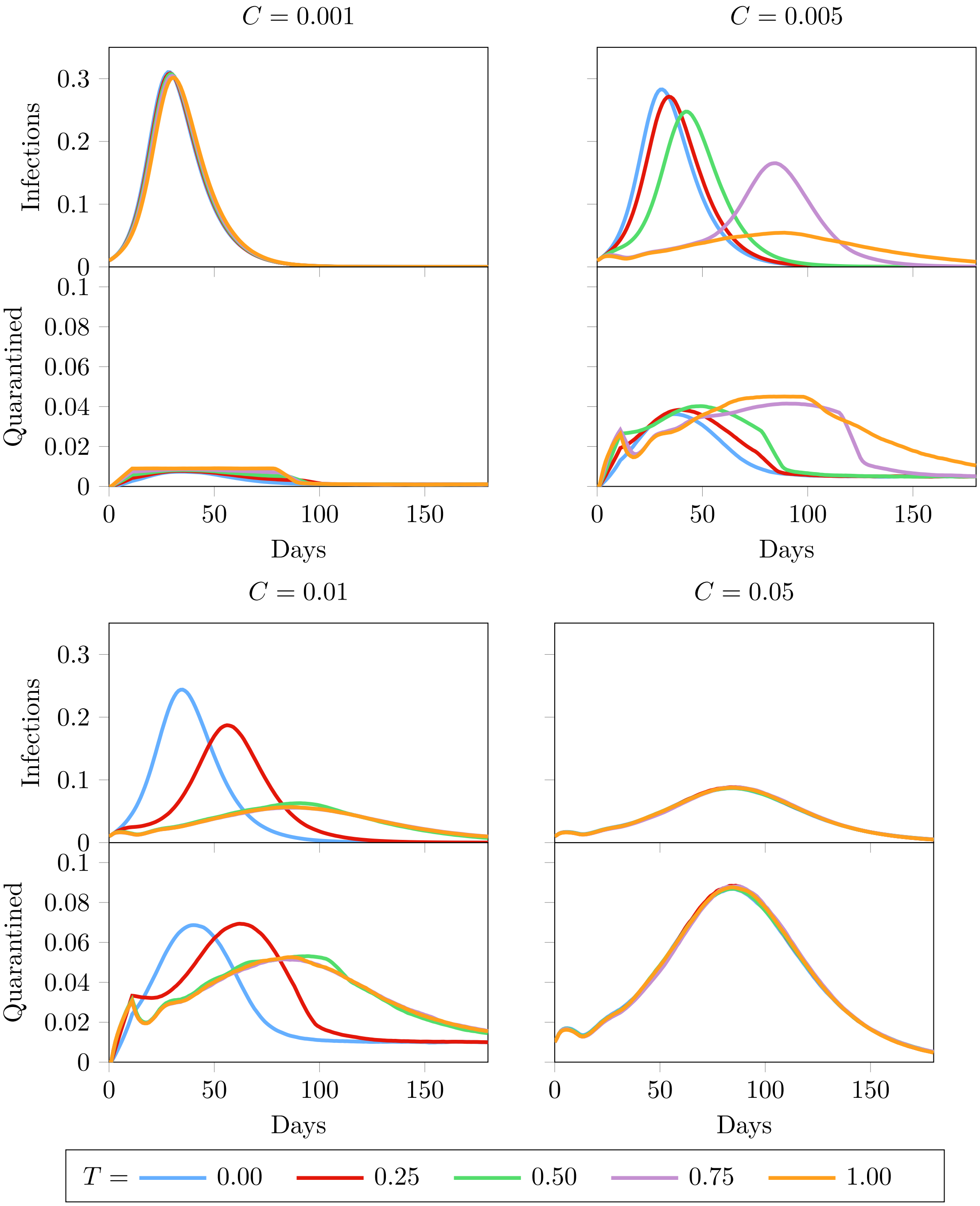


Figure 6.7: A comparison of different test targeting strategies $T$. Model parameters shown in Table 6.2.

know their viral history, and hence would not know which state they begin in. The model evaluations explore a range of capacities, specificities and seroprevalences for the antibody testing. These capacities and specificities, along with the capacity for testing, govern the transition of individuals from $\mathcal{QR} \longrightarrow \mathcal{R}$ (true positive tests), and from $\mathcal{QS} \longrightarrow \mathcal{S}$ (false positive tests). The sensitivity of the antibody test has little impact on the disease dynamics.

Figures 6.8 and 6.9 show the results of the model evaluations comparing the testing capacities and the specificities of the antibody test, the parameters for these runs are shown in Tables 6.3 and 6.4. The top row of each figure corresponds to the number of infections in time, the bottom row is the proportion of the population that are released from quarantine and hence are now in the working population. The steps in the plot with $C = 0.05$ are due to the model reaching a point at which everyone who wants a test can gets a test, and therefore there is a waiting period before more people become eligible.

Ultimately, the size of the subsequent peak depends on the rate at which people leave quarantine. In the plots with $C = 0.001$ and $C = 0.005$, with $t = 0.98$ the number of infections eventually decreases to zero, however, this is ultimately the impact of very few people being released. The models with $C = 0.001$ and $C = 0.005$, demonstrate that it is possible to keep the number of infections relatively stable for long periods of time, which may be beneficial if that level does not overwhelm healthcare services. Maximising the rate of reentry into the population is of course desirable, and it is widely appreciated that some increase in the numbers of infections is unavoidable. The desirable threshold in the trade-off between societal activity and number of infections is open to debate.

Each of the plots in Figure 6.9, show the effect of different seroprevalence in the population. To be clear, this is the proportion of the population that has contracted the virus and recovered but are in quarantine. The analysis has explored a range of seroprevalence from 0.05 to 0.95. These plots indicate that at high seroprevalence there is little benefit to better performing tests. For tests with high specificity the effect of the specificity is reduced. Once again the rates at which people leave quarantine has the largest impact on the subsequent size of the waves.

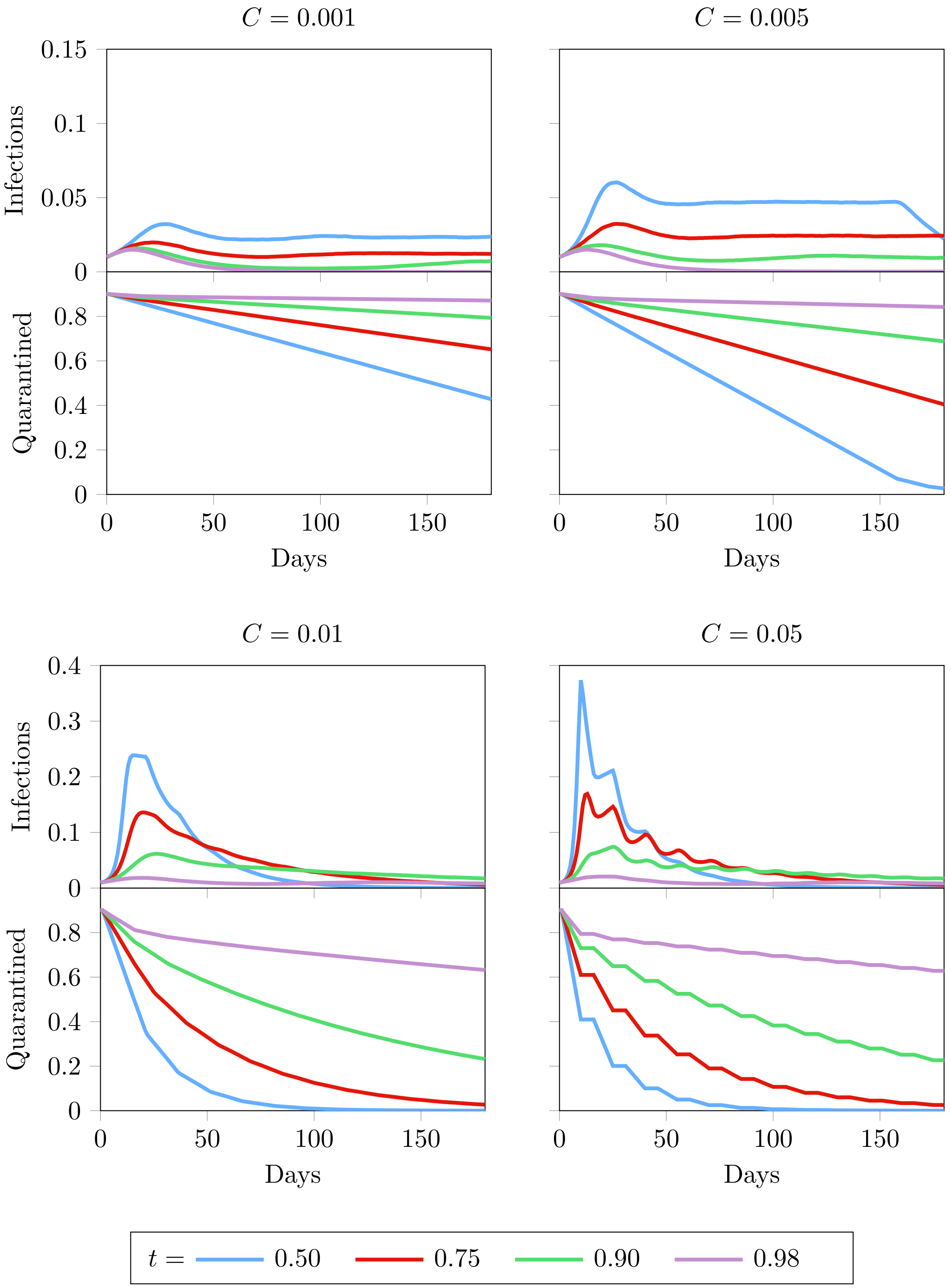


Figure 6.8: A comparison of different antibody test specificities with varying different test capacities. Top: The number of infected individuals ($\mathcal{I} + \mathcal{QI}$ population) over one year. Bottom: The proportion of the population that has been released from quarantine ($\mathcal{S} + \mathcal{I} + \mathcal{R}$ population) over one year. Model parameters are shown in Table 6.3.

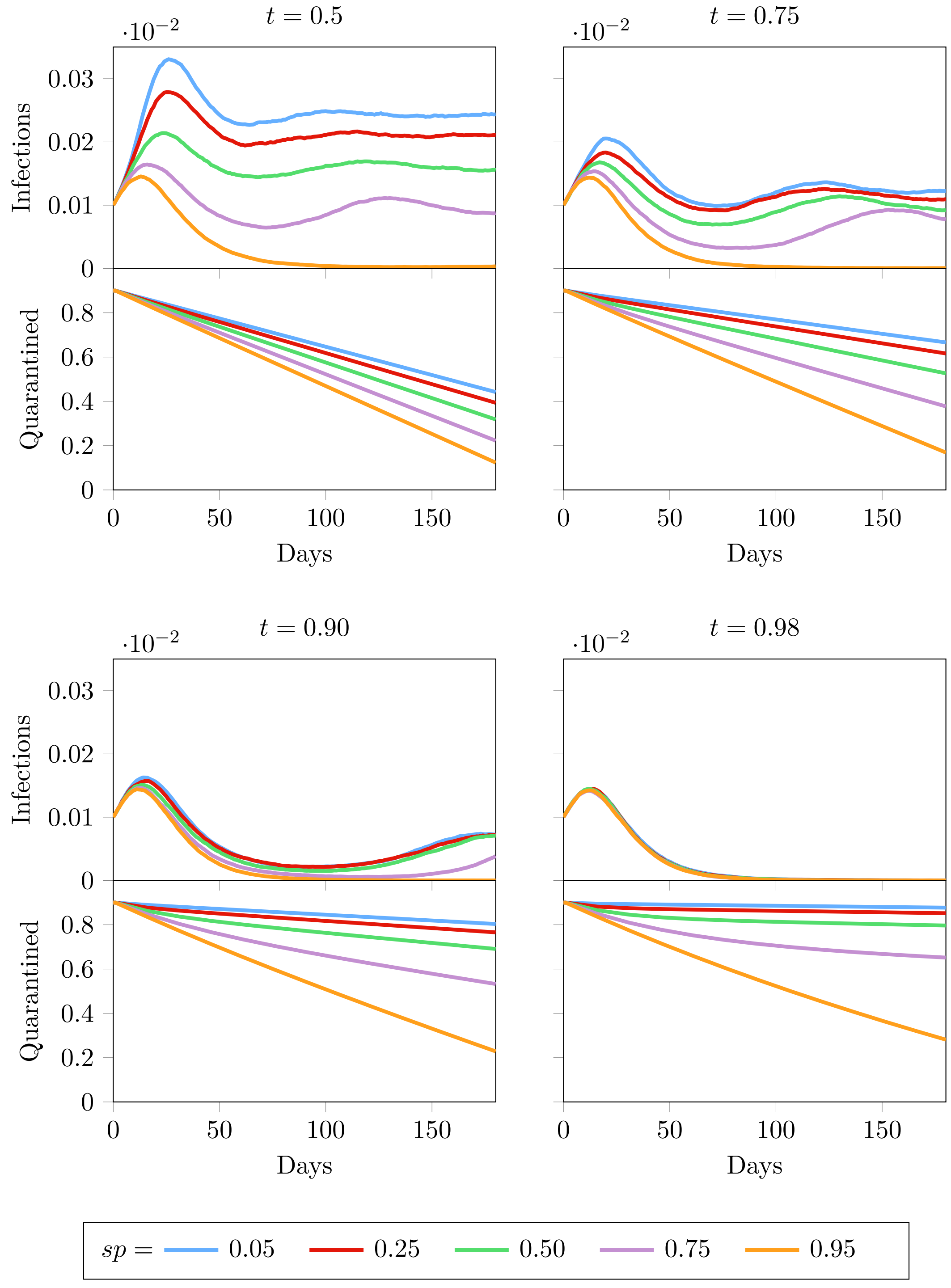


Figure 6.9: A comparison of different antibody test specificities with varying levels of seroprevalence ($sp$). Top: The number of infected individuals ($\mathcal{I} + \mathcal{QI}$ population) over one year. Bottom: The proportion of the population that has been released from quarantine ($\mathcal{S} + \mathcal{I} + \mathcal{R}$ population) over one year. Model parameters are shown in Table 6.4.

| Epidemic Parameters | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\beta = 0.32$ | | | | $\gamma = 0.1$ | |
| Antibody Test Parameters | | | | | | | |
| $C$ | $s$ | $t$ | $\Pi$ | $\Psi$ | $X_+$ | $X_-$ | $T$ |
| varies | 0.9 | varies | 1 | 1 | $\infty$ | 10 | 0 |
| Initial Population split | | | | | | | |
| N | $\mathcal{S}$ | $\mathcal{I}$ | $\mathcal{R}$ | $\mathcal{QS}$ | $\mathcal{QI}$ | $\mathcal{QR}$ | |
| $10^6$ | 0.5 | 0.01 | 0.04 | 0.8 | 0 | 0.1 | |

Table 6.3: Fixed parameters used for Figure 6.8 analysis. Antibody tests were disabled for this analysis.

| Epidemic Parameters | | | | | | | |
|---|---|---|---|---|---|---|---|
| | $\beta = 0.32$ | | | | | $\gamma = 0.1$ | |
| Antibody Test Parameters | | | | | | | |
| $C$ | $s$ | $t$ | $\Pi$ | $\Psi$ | $X_+$ | $X_-$ | $T$ |
| 0.005 | 0.9 | varies | 1 | 1 | $\infty$ | 10 | 0 |
| Initial Population split | | | | | | | |
| N | $\mathcal{S}$ | $\mathcal{I}$ | $\mathcal{R}$ | $\mathcal{QS}$ | $\mathcal{QI}$ | $\mathcal{QR}$ | |
| $10^6$ | 0.05 | 0.01 | 0.04 | $0.9(1-sp)$ | 0 | $0.9sp$ | |

Table 6.4: Fixed parameters used for Figure 6.9 analysis. Infection tests were disabled for this analysis. $sp$ is the seroprevalence of antibodies in the population.

## 6.4 Discussion

This analysis supports the assertion that a bad test is potentially worse than no test, but a good test is only effective in a carefully designed strategy. More is not necessarily better and over estimation of the test accuracy could be extremely detrimental.

This analysis is not a prediction; the numbers used are estimates and the SIRQ model is unlikely to be detailed enough to inform policy decisions. As such, the the analysis presented is not suited for drawing firm conclusions about the absolute necessary capacity of tests, nor the necessary sensitivity or specificity of tests or the recommended rate of release from quarantine. The analysis does, however, propose some conclusions that would broadly apply when testing and quarantining regimes are used to suppress epidemics:

- Diagnostic uncertainty can have a large effect on the dynamics of an epidemic, and sensitivity, specificity, and the capacity for testing alone are not sufficient to design effective testing procedures. Policy makers need to be aware of the accuracy of the tests, the prevelence of the disease at increased granularity and the characteristics of the target population when deciding on testing strategies.

- For infection screening to be used to relax quarantine measures the capacity needs to be sufficiently large but also well targeted to be effective. For example, this could be achieved through effective contact tracing.

- Mass screenings at large capacities may negate some of the need to effectively target tests. This occurs when the testing capacity is larger than the number of people who want to be tested, thus making the capacity de facto unlimited. This situatio n is broadly what occured within the UK in 2021.

- Caution should be exercised in the use of antibody testing. Assuming that the prevalence of antibodies is low, it is unlikely antibody testing at any scale will support the end of lockdown measures. And, untargeted antibody screening at population level could cause more harm than good.

- Antibody testing with a high specificity may be useful on an individual basis and has scientific value, which could reduce risk for key workers. But any belief that these tests would be useful to relax lockdown measures for the majority of the population is misguided.

- The analysis of antibody-test-based immunity passports would be broadly the same as those for vaccination-based immunity passports.

- As the prevelence of the disease is suppressed in different regions, it may be the case that small spikes in cases could be the result of false positives. This problem is potentially exacerbated by increased testing in localities in response to small increases in positive tests. Policy decisions that depend on small changes in the number of positive tests may, therefore, be flawed.

Epidemiological models used for policy making in real time will need to take into account the impact of diagnostic uncertainty of testing. The dynamical behaviour of modelled parameters in an appropriately complex model would need to include: quarantining; contact tracing and other surveillance strategies; test availability and targeting; and multiple sub-populations of susceptible, infected and recovered categories.

# Conclusions

This thesis has explored problems associated with enabling algorithms to compute with imprecise knowledge. This has been motivated by the discussion in Chapter 1 about issues related to an ever-increasing presence of algorithms within daily life. In particular, when faced with high-risk decisions, it is crucial to compute what is known rather than make assumptions that could have adverse outcomes leading from annoyance to injustice to catastrophe.

Chapter 2 introduced probability bounds analysis as a way of computing with epistemic and aleatory uncertainties. Intervals, p-boxes and c-boxes are explored, and example calculations were used to demonstrate how algorithms can use these objects to propagate both types of uncertainty.

Chapter 3 discussed the possibility of creating an automatic uncertainty compiler that would be able to introduce probability bounds analysis as a way of add uncertainty into pre-existing code. The chapter explored the components required to create such a compiler and discussed the difficulties associated with creating such a compiler. Whilst it is unlikely to be the case that such a compiler could be perfectly realised, a software tool that is able to help analysts who are unable to introduce uncertainty analysis into their codes would be extremely beneficial.

Chapter 4 discussed ways of correcting the performance statistics for binary classification tests when there is an imperfect gold standard. Several approaches were explored, including a novel method based on confidence boxes. However, it was concluded that finding a binary classifier's true performance is difficult when the ground truth is uncertain.

Chapter 5 considered the problem of imprecise information in logistic regression models. It was shown that it is possible to include interval uncertainty in both outcome status and predictor variables within logistic regression analysis by considering the set of possible models as an imprecise structure. Such a method expresses the epistemic uncertainties within the dataset that traditional methods have neglected. The imprecise bounds have the advantage of always containing the true model even when the data has been intervalised in biased or non-random ways. The outputs of the imprecise model prevent overly confident decisions being made where the confidence is a result of neglecting the imprecision. This has the potential to improve decisions in high risk situations.

Chapter 6 discussed how imperfect COVID tests may have impacted the spread of disease. It was shown how imprecision within a testing programme can affect the number of infections and the number of people quarantined, and that the sensitivity, specificity and how well the tests are targeted have significant effects. It is possible that including and communicating this imprecision may have lead to different public health decisions during the COVID-19 pandemic.

Ultimately, the tyranny of inhumane algorithms will be a significant risk for the foreseeable future. This thesis has argued that better handing of uncertainty will be an important part of understanding and mitigating this risk. This work is offered as practical steps towards to the aim of creating humane algorithms.

# List of Figures

# List of Tables

# List of Algorithms

# References

[1] N. Gray, D. Calleja, A. Wimbush, E. Miralles-Dolz, A. Gray, M. De Angelis, E. Derrer-Merk, B.U. Oparaji, V. Stepanov, L. Clearkin, and S. Ferson. Is “no test better than a bad test”? Impact of diagnostic uncertainty in mass testing on the spread of COVID-19. *PLoS ONE*, 15(10), 2020. doi: 10.1371/journal.pone.0240775.

[2] Nicholas Gray, Scott Ferson, Marco De Angelis, Ander Gray, and Francis Baumont de Oliveira. Probability bounds analysis for Python. *Software Impacts*, 12:100246, May 2022. ISSN 26659638. doi: 10.1016/j.simpa.2022.100246.

[3] Nicholas Gray, Marco De Angelis, and Scott Ferson. Towards an Automatic Uncertainty Compiler. *International Journal of Approximate Reasoning*, 2023 (*In Press*).

[4] Nick Gray, Marco De Angelis, Dominic Calleja, and Scott Ferson. A Problem in the Bayesian Analysis of Data without Gold Standards. In *29th European Safety and Reliability Conference*, pages 2628–2634, Hanover, Germany, 2019.

[5] Nick Gray, Marco De Angelis, and Scott Ferson. Computing With Uncertainty: Introducing Puffin the Automatic Uncertainty Compiler. In V. Papadopoulos M. Papadrakakis, G. Stefanou, editor, *3rd ECCOMAS Thematic Conference on Uncertainty Quantification in Computational Sciences and Engineering*, pages 487–497, Heraklion, Greece, 2019. doi: 10.7712/120219.6354.18702.

[6] Probability Bounds Analysis for Python. https://pypi.org/project/pba/{or}https://github.com/Institute-for-Risk-and-Uncertainty/pba-for-python/, .

[7] Alexander Wimbush, Nicholas Gray, and Scott Ferson. Singhing with confidence: Visualising the performance of confidence procedures. *Journal of Statistical Computation and Simulation*, pages 1–17, March 2022. ISSN 0094-9655, 1563-5163. doi: 10.1080/00949655.2022.2044814.

[8] Francis J. Baumont de Oliveira, Scott Ferson, Ronald A. D. Dyer, Jens M. H. Thomas, Paul D. Myers, and Nicholas G. Gray. How High Is High Enough? Assessing Financial

Risk for Vertical Farms Using Imprecise Probability. *Sustainability*, 14(9):5676, May 2022. ISSN 2071-1050. doi: 10.3390/su14095676.

[9] Nicholas Gray, Marco De Angelis, and Scott Ferson. The Creation of Puffin, the Automatic Uncertainty Compiler. *arXiv:2110.10153 [cs, stat]*, October 2021. http://arxiv.org/abs/2110.10153.

[10] Nicholas Gray and Scott Ferson. Logistic Regression Through the Veil of Imprecise Data. *arXiv:2106.00492 [stat]*, June 2021. http://arxiv.org/abs/2106.00492.

[11] Nicholas Diakopoulos. Algorithmic Accountability Reporting: On the Investigation of Black Boxes. Technical report, Columbia University, New York, NY, USA, 2014.

[12] Tom Scott. There is No Algorithm for Truth. https://youtu.be/leX541Dr2rU, September 2019.

[13] Isaac Asimov. *I, Robot*. Doubleday, London, UK, 1950. ISBN 978-0-00-827955-4.

[14] Susan Leigh Anderson. The Unacceptability of Asimov's Three Laws of Robotics as a Basis for Machine Ethics. In Michael Anderson and Susan Leigh Anderson, editors, *Machine Ethics*, pages 285–296. Cambridge University Press, Cambridge, 2011. ISBN 978-0-511-97803-6. doi: 10.1017/CBO9780511978036.021.

[15] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151(April 2018):90–103, 2019. doi: 10.1016/j.obhdp.2018.12.005.

[16] Esther Kaufmann. Algorithm appreciation or aversion? Comparing in-service and pre-service teachers' acceptance of computerized expert models. *Computers and Education: Artificial Intelligence*, 2:100028, 2021. ISSN 2666920X. doi: 10.1016/j.caeai.2021.100028.

[17] Johannes Schwienbacher. Reactions on Algorithms: A systematic Literature Review of Algorithm Aversion and Algorithm Appreciation. Master's thesis, University of Innsbruck, Austria, 2020.

[18] Yoyo Tsung-Yu Hou and Malte F. Jung. Who is the Expert? Reconciling Algorithm Aversion and Algorithm Appreciation in AI-Supported Decision Making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–25, October 2021. ISSN 2573-0142. doi: 10.1145/3479864.

[19] Hannah Fry. *Hello World: How to Be Human in the Age of the Machine*. WW Norton & Company Inc, New York, NY, USA, 2018.

[20] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, September 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0088-2.

[21] Charles Babbage. Difference Engine No. 1. In *Passages from the Life of a Philosopher*, pages 41–67. Longman, Green, Longman, Roberts and Green, London, 1864. ISBN 978-1-139-10367-1. doi: 10.1017/CBO9781139103671.006.

[22] George H. Lockwood. Final report of the Board of Inquiry investigating the circumstances of an accident involving the Air Canada Boeing 767 aircraft C-GAUN that effected an emergency landing at Gimli, Manitoba on the 23rd day of July, 1983. Technical Report 066011884X, Ministry of Transport, Ottawa, Canada, 1985.

[23] A.G. Stephenson, D.R. Mulville, F.H. Bauer, G.A. Dukeman, P. Norvig, L.S. LaPiana, and R. Sackheim. Mars Climate Orbiter Mishap Investigation Board Phase I Report. *National Aeronautics and Space Administration*, pages 44–44, 1999. ISSN 9781450344036. doi: 10.1145/3059454.3059463.

[24] Bridget M Kuehn. Group Urges Going Metric to Head Off Dosing Mistakes. *The Journal of the American Medical Association*, 311(21):2159–2160, 2015.

[25] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3): 457–506, March 2021. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-021-05946-3.

[26] Scott Ferson, Jason O'Rawe, Andrei Antonenko, Jack Siegrist, James Mickley, Christian C. Luhmann, Kari Sentz, and Adam M. Finkel. Natural language of uncertainty: Numeric hedge words. *International Journal of Approximate Reasoning*, 57:19–39, February 2015. doi: 10.1016/J.IJAR.2014.11.003.

[27] Sara Baase. *A Gift of Fire*. Pearson Education, Upper Saddle River, NJ, USA, second edition, 2003.

[28] Nancy G. Leveson and Clark S. Turner. An Investigation of the Therac-25 Accidents. *IEEE Computer*, 26(7):18–41, 1993.

[29] Nancy G. Leveson. Appendix: Medical Devices: The Therac-25. In *Software: Safety Systems and Computers*. 1995.

[30] Nancy G. Leveson. The Therac-25: 30 Years Later. *Computer*, 50(11):8–11, November 2017. ISSN 0018-9162. doi: 10.1109/MC.2017.4041349.

[31] Harold Thimbleby. Is IT a dangerous prescription? *BCS Interfaces*, 84:5–10, 2010.

[32] H. Thimbleby and P. Cairns. Reducing number entry errors: Solving a widespread, serious problem. *Journal of the Royal Society Interface*, 7(51):1429–1439, 2010. doi.org/10.1098/rsif.2010.0112.

[33] BEA. Final Report On the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight AF 447 Rio de Janeiro - Paris (English Language Version). Technical Report 8811117690, Ministère de l'Écologie, du Développement durable, des Transports et du Logement, Paris, France, 2012.

[34] SKYbrary. Unreliable Airspeed Indications. https://skybrary.aero/articles/unreliable-airspeed-indications, 2022.

[35] Natalia Criado and Jose M Such. *Digital Discrimination.* Oxford University Press, first edition, September 2019. ISBN 978-0-19-883849-4 978-0-19-187472-7. doi: 10.1093/oso/9780198838494.001.0001.

[36] Natalia Criado, Xavier Ferrer-Aran, and Jose M Such. Is my program sexist? Using Norms to Attest Digital Discrimination. *IEEE Technology and Society Magazine*, page 8, 2020.

[37] BBC News. Apple's 'sexist' credit card investigated by US regulator. *BBC News*, November 2019. https://www.bbc.com/news/business-50365609.

[38] Jesselyn Cook. Instagram's Shadow Ban On Vaguely 'Inappropriate' Content Is Plainly Sexist. https://www.huffpost.com/entry/instagram-shadow-ban-sexist_n_5cc72935e4b0537911491a4f, April 2019.

[39] Jesselyn Cook. Women Are Pretending To Be Men On Instagram To Avoid Sexist Censorship. https://www.huffingtonpost.com.au/entry/women-pretending-to-be-men-instagram-censorship_au_5dd4532ce4b0fc53f20a140d, November 2019.

[40] Andres Ferraro, Xavier Serra, and Christine Bauer. Break the Loop: Gender Imbalance in Music Recommenders. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pages 249–254, Canberra ACT Australia, March 2021. ACM. ISBN 978-1-4503-8055-3. doi: 10.1145/3406522.3446033.

[41] Caroline Criado-Perez. *Invisible Women: Exposing Data Bias in a World Designed for Men.* Chatto & Windus, London, Uk, 2019. ISBN 978-1-78474-172-3 978-1-78474-292-8.

[42] Latanya Sweeney. Discrimination in Online Ad Delivery: Google ads, black names

and white names, racial discrimination, and click advertising. *Queue*, 11(3):10–29, March 2013. ISSN 1542-7730, 1542-7749. doi: 10.1145/2460276.2460278.

[43] Alex Hern. Twitter apologises for 'racist' image-cropping algorithm. *The Guardian*, September 2020. ISSN 0261-3077. https://www.theguardian.com/technology/2020/sep/21/twitter-apologises-for-racist-image-cropping-algorithm.

[44] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias. *ProPublica*, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[45] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How We Analyzed the COMPAS Recidivism Algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm, 2016.

[46] Anon. Sample Risk Assessment COMPAS. https://s3.documentcloud.org/documents/2702103/Sample-Risk-Assessment-COMPAS-CORE.pdf, 2016.

[47] John J. Donohue III and Steven D. Levitt. The Impact of Race on Policing and Arrests. *The Journal of Law and Economics*, 44(2):367–394, October 2001. ISSN 0022-2186, 1537-5285. doi: 10.1086/322810.

[48] Richard A. Berk. Artificial Intelligence, Predictive Policing, and Risk Assessment for Law Enforcement. *Annual Review of Criminology*, 4(1):209–237, January 2021. ISSN 2572-4568, 2572-4568. doi: 10.1146/annurev-criminol-051520-012342.

[49] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. fairmlbook.org.

[50] Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, April 2020. ISSN 0001-0782, 1557-7317. doi: 10.1145/3376898.

[51] Anthony W Flores, Kristin Bechtel, and Christopher T. Lowenkamp. False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks.". *Federal Probation*, 80(2):10, 2016.

[52] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):1–6, 2018. doi: 10.1126/sciadv.aao5580.

[53] John Tyler Clemons. Blind Injustice: The Supreme Court, Implicit Racial Bias, and the Racial Disparity in the Criminal Justice System. *American Criminal Law Review*, 51:27, 2014.

[54] S. Danziger, J. Levav, and L. Avnaim-Pesso. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892, April 2011. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1018033108.

[55] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General*, 144(1):114–126, 2014.

[56] Kristian Lum and William Isaac. To predict and serve? *Significance*, 13(5):14–19, 2016. ISSN 1740-9713. doi: 10.1111/j.1740-9713.2016.00960.x.

[57] Bryce Goodman and Seth Flaxman. European Union regulations on algorithmic decision-making and a “right to explanation”. *AI Magazine*, 38(3):1–9, 2016.

[58] Michael Scott Balch, Ryan Martin, and Scott Ferson. Satellite conjunction analysis and the false confidence theorem. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 475(20180565), 2019. doi: 10.1098/rspa.2018.0565.

[59] Hitaf R. Kady and Jennifer A. Vadeboncoeur, PhD. Digital divide. *Salem Press Encyclopedia*, 2019. https://liverpool.idm.oclc.org/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=ers&AN=89677542&site=eds-live&scope=site.

[60] Anne-Britt Gran, Peter Booth, and Taina Bucher. To be or not to be algorithm aware: A question of a new digital divide? *Information, Communication & Society*, 24(12): 1779–1796, September 2021. ISSN 1369-118X, 1468-4462. doi: 10.1080/1369118X.2020.1736124.

[61] Turner Osler, Laurent G Glance, and David W Hosmer. Simplified Estimates of the Probability of Death After Burn Injuries : Extending and Updating the Baux Score. *Journal of Trauma Injury, Infection and Critical Care*, 68(3), 2010. doi: 10.1097/TA.0b013e3181c453b3.

[62] David W. Hosmer Jr, Stanley Lameshow, and Rodney X. Sturdivant. *Applied Logistic Regression*. John Wiley & Sons, Ltd, Hoboken, NJ, USA, 3rd edition, 2013.

[63] Miranda Fricker. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford Scholarship Online, 2007. ISBN 978-0-19-823790-7. doi: 10.1093/acprof:oso/9780198237907.001.0001.

[64] Joel Walmsley. Artificial intelligence and the value of transparency. *AI & Society*, September 2020. ISSN 0951-5666, 1435-5655. doi: 10.1007/s00146-020-01066-z.

[65] GDPR. Regulation (EU) 2016/679 on the protection of natural persons with regard to

the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (Data Protection Directive). OJ L 119/1, 2016.

[66] Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv*, (MI):1–13, 2017.

[67] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics*, Turin, Italy, 2018.

[68] Lilian Edwards and Michael Veale. Slave to the Algorithm? Why a right to an explanation is probably not the remedy you are looking for. *Duke Law & Technology Review*, 16(1):15–84, 2017. ISSN 3540445668.

[69] Mohamed Jehad Baeth and Mehmet S Aktas. Detecting misinformation in social networks using provenance data. *Concurrency and Computation: Practice and Experience*, 31(3):13, 2018.

[70] Lilian Edwards. Personal Communication, June 2022.

[71] Center for Countering Digital Hate. Failure to Act: How Tech Giants Continue to Defy Calls to Rein in Vaccine Misinformation. Technical report, Washington, DC, USA and London, UK, 2020.

[72] Center for Countering Digital Hate. The Disinformation Dozen. Technical report, Washington, DC, USA and London, UK, 2021.

[73] Center for Countering Digital Hate. The Toxic Ten: How ten fringe publishers fuel 69% of digital climate change denial. Technical report, Center for Countering Digital Hate, Washington, DC, USA and London, UK, 2021.

[74] Mike Shafto, Mike Conroy, Rich Doyle, Ed Glaessgen, Chris Kemp, Jacqueline LeMoigne, and Lui Wang. Modeling, Simulation, Information Technology and Processing Roadmap - Technology Area 11. Technical report, National Air and Space Administration, Washington, DC, USA, 2012.

[75] Stefan Boschert and Roland Rosen. Digital Twin—The Simulation Aspect. In *Mechatronic Futures*, pages 59–74. Springer International Publishing, Cham, 2016. ISBN 978-3-319-32156-1. doi: 10.1007/978-3-319-32156-1.

[76] Juuso Autiosalo. Platform for industrial internet and digital twin focused education, research, and innovation: Ilmatar the overhead crane. In *IEEE World Forum on*

*Internet of Things, WF-IoT 2018 - Proceedings*, volume 2018-Janua, pages 241–244, 2018. ISBN 978-1-4673-9944-9. doi: 10.1109/WF-IoT.2018.8355217.

[77] David Jones, Chris Snider, Aydin Nassehi, Jason Yon, and Ben Hicks. Characterising the Digital Twin: A systematic literature review. *CIRP Journal of Manufacturing Science and Technology*, 29:36–52, May 2020. ISSN 17555817. doi: 10.1016/j.cirpj.2020.02.002.

[78] K. Worden, E. J. Cross, P. Gardner, R. J. Barthorpe, and D. J. Wagg. On Digital Twins, Mirrors and Virtualisations. In Robert Barthorpe, editor, *Model Validation and Uncertainty Quantification, Volume 3*, pages 285–295. Springer International Publishing, Cham, 2019. ISBN 978-3-030-12074-0 978-3-030-12075-7. doi: 10.1007/978-3-030-12075-7_34.

[79] Murat Dundar, Balaji Krishnapuram, Jinbo Bi, and R. Bharat Rao. Learning Classifiers When the Training Data Is Not IID. In *IJCAI'07*, page 6, 2007.

[80] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, and the CAMELYON16 Consortium, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, Oscar Geessink, Nikolaos Stathonikos, Marcory CRF van Dijk, Peter Bult, Francisco Beca, Andrew H Beck, Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, Aoxiao Zhong, Qi Dou, Quanzheng Li, Hao Chen, Huang-Jing Lin, Pheng-Ann Heng, Christian Haß, Elia Bruni, Quincy Wong, Ugur Halici, Mustafa Ümit Öner, Rengul Cetin-Atalay, Matt Berseth, Vitali Khvatkov, Alexei Vylegzhanin, Oren Kraus, Muhammad Shaban, Nasir Rajpoot, Ruqayya Awan, Korsuk Sirinukunwattana, Talha Qaiser, Yee-Wah Tsang, David Tellez, Jonas Annuscheit, Peter Hufnagl, Mira Valkonen, Kimmo Kartasalo, Leena Latonen, Pekka Ruusuvuori, Kaisa Liimatainen, Shadi Albarqouni, Bharti Mungal, Ami George, Stefanie Demirci, Nassir Navab, Seiryo Watanabe, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, Hady Ahmady Phoulady, Vassili Kovalev, Alexander Kalinovsky, Vitali Liauchuk, Gloria Bueno, M. Milagro Fernandez-Carrobles, Ismael Serrano, Oscar Deniz, Daniel Racoceanu, and Rui Venâncio. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, 318 (22):2199, December 2017. ISSN 0098-7484. doi: 10.1001/jama.2017.14585.

[81] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe

Thagaard, Anders B. Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih cetin, Eren Halici, Hunter Jackson, Richard Chen, Fabian Both, Jorg Franke, Heidi Kusters-Vandevelde, Willem Vreuls, Peter Bult, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens. From Detection of Individual Metastases to Classification of Lymph Node Status at the Patient Level: The CAMELYON17 Challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, February 2019. ISSN 0278-0062, 1558-254X. doi: 10.1109/TMI.2018.2867350.

[82] Silvia Moreno, Mario Bonfante, Eduardo Zurek, and Homero San Juan. Study of Medical Image Processing Techniques Applied to Lung Cancer. In *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6, Coimbra, Portugal, June 2019. IEEE. ISBN 978-989-98434-9-3. doi: 10.23919/CISTI.2019.8760888.

[83] Odette Wegwarth and Gerd Gigerenzer. *Statistical Illiteracy in Doctors*, page 16. MIT Press, Cambridge, MA, USA, 2011.

[84] Gerd Gigerenzer and Ulrich Hoffrage. How to Improve Bayesian Reasoning Without Instruction: Frequency Formats. *Psychological Review,*, 102(4):684–704, 1995.

[85] Gerd Gigerenzer. What are natural frequencies? *BMJ (Online)*, 343(7828):1–2, 2011. doi: 10.1136/bmj.d6386.

[86] Mirta Galesic, Rocio Garcia-Retamero, and Gerd Gigerenzer. Using icon arrays to communicate medical risks: Overcoming low numeracy. *Health Psychology*, 28(2):210–216, 2009. ISSN 1930-7810, 0278-6133. doi: 10.1037/a0014474.

[87] Brian J. Zikmund-Fisher, Holly O. Witteman, Mark Dickson, Andrea Fuhrel-Forbis, Valerie C. Kahn, Nicole L. Exe, Melissa Valerio, Lisa G. Holtzman, Laura D. Scherer, and Angela Fagerlin. Blocks, Ovals, or People? Icon Type Affects Risk Perceptions and Recall of Pictographs. *Medical Decision Making*, 34(4):443–453, May 2014. ISSN 0272-989X, 1552-681X. doi: 10.1177/0272989X13511706.

[88] Becky Woods. Alton Towers Smiler ride reopens nine months after horror crash. *BBC News Online*, March 2016. https://www.bbc.co.uk/news/uk-england-stoke-staffordshire-35851291.

[89] Stephen Flanagan. Re: Accident at Smiler Rollercoaster, Alton Towers, 2nd June 2015. Expert's Report, October 2015. Technical report, Health and Safety Executive (United Kingdom), 2015. https://www.whatdotheyknow.com/request/alton_towers_smiler_crash_report.

[90] Christopher A. Hart. Boeing 737 MAX Flight Control System Observations, Findings and Recommendations. Technical report, U.S. Federal Aviation Administration, Washington, DC, USA, 2019. https://www.faa.gov/news/media/attachments/Final_JATR_Submittal_to_FAA_Oct_2019.pdf.

[91] KNKT. Final Aircraft Accident Investigation Report PT. Lion Mentari Airlines Boeing 737-8(MAX); PK-LQP Tanjung Karawang, West Java Republic of Indonesia 29 October 2018. Technical report, Komite Nasional Keselamatan Transportasi, Jakarta, Indonesia, 2019. http://knkt.dephub.go.id/knkt/ntsc_home/ntsc.htm.

[92] ECAA. Aircraft Accident Investigation Preliminary Report: B737-8 (MAX) Registered ET-AVJ 28 NM South East of Addis Ababa, Bole International Airport. Technical report, Ethiopian Civil Aviation Authority, Ministry of Transport, 2019. http://www.ecaa.gov.et/documents/20435/0/Preliminary+Report+B737-800MAX+,(ET-AVJ).pdf.

[93] Michael A. Cusumano. Boeing's 737 MAX: A failure of management, not just technology. *Communications of the ACM*, 64(1):22–25, January 2021. ISSN 0001-0782, 1557-7317. doi: 10.1145/3436231.

[94] Lisanne Bainbridge. Ironies of automation. *Automatica*, 19(6):775–779, 1983. doi: 10.1016/0005-1098(83)90046-8.

[95] SAE. The Principles of Operation Framework: A Comprehensive Classification Concept for Automated Driving Functions. *SAE International Journal of Connected and Automated Vehicles*, 3(1):12–03–01–0003, 2021. ISSN 2574-075X. doi: 10.4271/12-03-01-0003.

[96] NTSB. Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian, Tempe, Arizona, March 18, 2018. Technical report, National Transportation Safety Board, Washington, DC, USA, 2019.

[97] Phillipa Foot. The Problem of Abortion and the Doctrine of the Double Effect. *Oxford Review*, 5, 1967.

[98] Judith J. Thomson. Killing, Letting Die and the Trolley Problem. *Monist: An International Quaterly Journal of General Philisophical Inquiry*, 59:204–217, 1976.

[99] Judith J. Thomson. The Trolley Problem. *The Yale Law Journal*, 94(6):1935–1415, 1985.

[100] F. M. Kamm. Harming Some to Save Others. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 57(3):227–260, 1989.

[101] Peter Unger. *Living High and Letting Die*. Oxford Universithy Press, New York, NY, USA, 1996.

[102] Barbara H. Fried. What does matter? the case for killing the trolley problem (or letting it die). *Philosophical Quarterly*, 62(248):505–529, 2012. doi: 10.1111/j.1467-9213.2012.00061.x.

[103] Sven Nyholm and Jilles Smids. The Ethics of Accident-Algorithms for Self-Driving Cars: An Applied Trolley Problem? *Ethical Theory and Moral Practice*, 19(5):1275–1289, November 2016. doi: 10.1007/s10677-016-9745-2.

[104] Johannes Himmelreich. Never Mind the Trolley: The Ethics of Autonomous Vehicles in Mundane Situations. *Ethical Theory and Moral Practice*, 21(3):669–684, 2018. doi: 10.1007/s10677-018-9896-4.

[105] Lord Justice Green, Sarah Green, Nick Hopkins, Penney Lewis, Nicholas Paines, Lady Paton, David Bartos, Gillian Black, Kate Dowdalls, and Frankie McCarthy. Automated Vehicles: Joint report. Technical report, Law Commission of England and Wales and Scottish Law Commission, London, UK, 2022.

[106] Adam M. Finkel. Protecting People in Spite of–or Thanks to–the “Veil of Ignorance”. In Richard R. Sharp and Gary E. Marchant, editors, *Genomics and Environmental Regulation: Science, Ethics, and Law*, pages 290–342. The John Hopkins University Press, Balitmore, 2008.

[107] Australian Bureau of Statistics. 4102.0 - Australian Social Trends, April 2013. Technical report, Australian Bureau of Statistics, 2013. https://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/4102.0Main+Features30April+2013.

[108] Mary Mallappallil, Jacob Sabu, Angelika Gruessner, and Moro Salifu. A review of big data and medical research. *SAGE Open Medicine*, 8:205031212093483, January 2020. ISSN 2050-3121, 2050-3121. doi: 10.1177/2050312120934839.

[109] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, April 2015. ISSN 2299-0984. doi: 10.1515/popets-2015-0007.

[110] Hal Berghel. Malice Domestic: The Cambridge Analytica Dystopia. *Computer*, 51 (5):84–89, May 2018. ISSN 0018-9162. doi: 10.1109/MC.2018.2381135.

[111] I. Hrynaszkiewicz, M. L Norton, A. J Vickers, and D. G Altman. Preparing raw clinical data for publication: Guidance for journal editors, authors, and peer reviewers. *BMJ*,

340(jan28 1):c181–c181, January 2010. ISSN 0959-8138, 1468-5833. doi: 10.1136/bmj.c181.

[112] Jules Polonetsky, Omer Tene, and Kelsey Finch. Shades of Gray: Seeing the Full Spectrum of Practical Data De-identification. *Santa Clara Law Review*, 56:39, 2016.

[113] Luc Rocher, Julien M. Hendrickx, and Yves Alexandre de Montjoye. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, 10(1), 2019. ISSN 4146701910. doi: 10.1038/s41467-019-10933-3.

[114] Philippe Golle. Revisiting the uniqueness of simple demographics in the US population. In *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society - WPES '06*, page 77, Alexandria, Virginia, USA, 2006. ACM Press. ISBN 978-1-59593-556-4. doi: 10.1145/1179601.1179615.

[115] Gang Xiang, Jason O Rawe, Vladik Krienovich, Janos Hajagos, and Scott Ferson. Protecting patient privacy while preserving medical information for research. In *Proceedings of the 6th International Workshop on Reliable Engineering Computing (REC'2014)*, pages 281–293, Chicago, Il, USA, 2014.

[116] Vladik Kreinovich and Christian Servin. How to Test Hypotheses When Exact Values Are Replaced by Intervals to Protect Privacy : Case of t-tests. *International Journal of Intelligent Technologies and Applied Statstics*, 8(2):93–102, 2015. doi: 10.6148/IJITAS.2015.0802.01.

[117] Luc Longpré, Vladik Kreinovich, and Thongchai Dumrongpokaphan. Entropy as a measure of average loss of privacy. *Thai Journal of Mathematics*, 15(Special Issue On Entropy In Econometrics):7–15, 2017.

[118] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, SciPy 1.0 Contributors, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm,

G. Young, Gavin A. Price, Gert-Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3): 261–272, March 2020. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-019-0686-2.

[119] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-020-2649-2.

[120] John D. Hunter. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. ISSN 1521-9615. doi: 10.1109/MCSE.2007.55.

[121] Scott Ferson. *RAMAS Risk Calc 4.0 Software: Risk Assessment with Uncertain Numbers*. Lewis Publishers, Boca Raton, Florida, USA, 2002. ISBN 978-1-56670-576-9. https://books.google.co.uk/books?id=tKz7UZRs0CEC.

[122] Probability Bounds Analysis for MATLAB. https://github.com/Institute-for-Risk-and-Uncertainty/pba-for-matlab, .

[123] Probability Bounds Analysis for R. https://github.com/ScottFerson/pba.r, .

[124] Probability Bounds Analysis for Julia. https://github.com/AnderGray/ProbabilityBoundsAnalysis.jl, .

[125] Ander Gray, Scott Ferson, and Edoardo Patelli. ProbabilityBoundsAnalysis.jl: Arithmetic with sets of distributions. In *JuliaCon*, page 12, Virtual Conference, 2021.

[126] Stack Overflow. 2021 Developer Survery. https://insights.stackoverflow.com/survey/2021, May 2021.

[127] TIOBE - The Software Quality Company. TIOBE Index. https://www.tiobe.com/tiobe-index/, January 2022.

[128] Christopher J. Roy and Michael S. Balch. A holistic approach to uncertainty quantification with application to supersonic nozzle thrust. *International Journal for Uncertainty Quantification*, 2(4):363–381, 2012. ISSN 2152-5080. doi: 10.1615/Int.J.UncertaintyQuantification.2012003562.

[129] Lloyd Goldwasser, Scott Ferson, and Lev Ginzburg. Variability and Measurement Error in Extinction Risk Analysis: The Northern Spotted Owl on the Olympic Peninsula. In *Quantitative Methods for Conservation Biology*, pages 169–187. Springer-Verlag, New York, 2000. ISBN 978-0-387-95486-8. doi: 10.1007/0-387-22648-6_11.

[130] Daniel J. Rozell and Sheldon J. Reaven. Water Pollution Risk Associated with Natural Gas Extraction from the Marcellus Shale: Marcellus Shale Water Pollution Risk. *Risk Analysis*, 32(8):1382–1393, August 2012. ISSN 02724332. doi: 10.1111/j.1539-6924.2011.01757.x.

[131] Luis G. Crespo, Sean P. Kenny, and Daniel P. Giesy. Reliability analysis of polynomial systems subject to p-box uncertainties. *Mechanical Systems and Signal Processing*, 37(1-2):121–136, May 2013. ISSN 08883270. doi: 10.1016/j.ymssp.2012.08.012.

[132] Michael Beer, Scott Ferson, and Vladik Kreinovich. Imprecise probabilities in engineering analyses. *Mechanical Systems and Signal Processing*, 37(1-2):4–29, 2013. doi: 10.1016/j.ymssp.2013.01.024.

[133] Jonathan Sadeghi, M. de Angelis, and Edoardo Patelli. Efficient training of interval Neural Networks for imprecise training data. *Neural Networks*, 118:338–351, 2019. doi: 10.1016/j.neunet.2019.07.005.

[134] Jonathan Cyrus Sadeghi. *Uncertainty Modelling for Scarce and Imprecise Data in Engineering Applications*. PhD thesis, University of Liverpool, 2020.

[135] Götz Alefeld and Günter Mayer. Interval analysis: Theory and applications. *Journal of Computational and Applied Mathematics*, 121(1):421–464, 2000. ISSN 0377-0427. doi: 10.1016/S0377-0427(00)00342-3.

[136] Federica Gioia, Carlo N Lauro, and Napoli Federico. Basic statistical methods for interval data. *Statistica Applicata*, 17:1–29, 2005.

[137] Timothy Hickey, Qun Ju, and Maarten H Van Emden. Interval Arithmetic: From Principles to Implementation. *Journal of the ACM*, 48(5):1038–1068, 2001. doi: 10.1145/502102.502106.

[138] IEEE. *1788-2015 – IEEE Standard for Interval Arithmetic*. 2015. ISBN 978-0-7381-9720-3. doi: 10.1109/IEEESTD.2015.7140721.

[139] Ramon E Moore, R Baker Kearfott, and Michael J Cloud. *Introduction to Interval Analysis*, volume 110. Society for Industrial and Applied Mathematics, Philadelphia, USA, 2009. ISBN 978-0-898716-6.

[140] Scott Ferson, Vladik Kreinovich, Lev Ginzburg, Davis S Myers, and Kari Sentz. Constructing Probability Boxes and Dempster-Shafer Structures. Technical Report January, Sandia National Laboratories, Albuquerque, NM, United States, 2003.

[141] Scott Ferson, Roger B Nelsen, Janos Hajagos, Daniel J Berleant, Jianzhong Zhang, W Troy Tucker, Lev R Ginzburg, and William L Oberkampf. Dependence in probabilistic modeling, Dempster-Shafer theory, and probability bounds analysis. Technical Report 19094, Sandia National Laboratories, Albuquerque, NM, USA, 2004.

[142] Scott Ferson and William L. Oberkampf. Validation of imprecise probability models. *International Journal of Reliability and Safety*, 3(1/2/3):3–3, 2009. doi: 10.1504/IJRS.2009.026832.

[143] William L Oberkampf and Scott Ferson. Model Validation under Both Aleatory and Epistemic Uncertainty. *New York*, (i):1–26, 2007. doi: 10.1007/s12200-008-0023-3.

[144] W. F. Mascarenhas. Moore: Interval Arithmetic in C++20. In Guilherme A. Barreto and Ricardo Coelho, editors, *37th Conference of the North American Fuzzy Information Processing Society*, volume 831, pages 519–529, Fortaleza, Brazil, 2018. Springer International Publishing. ISBN 978-3-319-95311-3 978-3-319-95312-0. doi: 10.1007/978-3-319-95312-0_45.

[145] HYRISK. https://cran.r-project.org/web/packages/HYRISK/index.html.

[146] Martin Haenggi. Meta Distributions–Part 1: Definition and Examples. *IEEE Communications Letters*, pages 1–1, 2021. ISSN 1089-7798, 1558-2558, 2373-7891. doi: 10.1109/LCOMM.2021.3069662.

[147] Martin Haenggi. Meta Distributions–Part 2: Properties and Interpretations. *IEEE Communications Letters*, pages 1–1, 2021. ISSN 1089-7798, 1558-2558, 2373-7891. doi: 10.1109/LCOMM.2021.3069681.

[148] J G Dijkman, H V A N Haeringen, and S J D E Lange. Fuzzy Numbers. *Journal of Mathematical Analysis and Applicatins*, 92:301–341, 1983.

[149] Didier Dubois and Henri Prade. Interval-valued Fuzzy Sets , Possibility Theory and Imprecise Probability. In *Proceedings of the Joint 4th Conference of the European Society for Fuzzy Logic and Technology and the 11th Rencontres Francophones Sur La Logique Floue et Ses Applications*, Barcelona, Spain, 2005.

[150] Michael Scott Balch. New two-sided confidence intervals for binomial inference derived using Walley's imprecise posterior likelihood as a test statistic. *International Journal of Approximate Reasoning*, 123:77–98, 2020. doi: 10.1016/j.ijar.2020.05.005.

[151] Yakob Ben-Haim. *Info-Gap Decision Theory: Decisions Under Severe Uncertainty*. Academic Press, Oxford, UK, second edition, 2006.

[152] Vladik Kreinovich. Interval Computations and Interval-Related Statistical Techniques: Tools for Estimating Uncertainty of the Results of Data Processing and Indirect Measurements. In Franco Pavese and Alistair B. Forbes, editors, *Data Modeling for Metrology and Testing in Measurement Science*, pages 1–29. Birkhäuser Boston, Boston, 2009. ISBN 978-0-8176-4592-2 978-0-8176-4804-6. doi: 10.1007/978-0-8176-4804-6_4.

[153] Luiz Henrique De Figueiredo and Jorge Stolfi. Affine arithmetic: Concepts and applications. *Numerical Algorithms*, 37:147–158, 2004.

[154] Eric Goubault and Sylvie Putot. A zonotopic framework for functional abstractions. *Formal Methods in System Design*, 47(3):302–360, 2016. doi: 10.1007/s10703-015-0238-z.

[155] Ander Gray, Marco De Angelis, Scott Ferson, and Edoardo Patelli. What's Z-X, when Z = X+Y? Dependency tracking in interval arithmetic with bivariate sets. In *9th International Workshop on Reliable Engineering Computations*, pages 27–28, Virtual Conference, 2021.

[156] Walter Krämer. Generalized Intervals and the Dependency Problem. *Proceedings in Applied Mathematics and Mechanics*, 684:683–684, 2006. doi: 10.1002/pamm.200610.

[157] Michael Scott Balch. Mathematical foundations for a theory of confidence structures. *International Journal of Approximate Reasoning*, 53(7):1003–1019, 2012. ISSN 0888-613X. doi: 10.1016/j.ijar.2012.05.006.

[158] Scott Ferson, Michael Balch, Kari Sentz, and Jack Siegrist. Computing with Confidence. In *Proceedings of the Eighth International Symposium on Imprecise Probability:*

*Theory and Applications*, Compiègne, France, 2013. http://www.sipta.org/isipta13/proceedings/papers/s013.pdf.

[159] Scott Ferson, Jason O'Rawe, and Michael Balch. Computing with Confidence: Imprecise Posteriors and Predictive Distributions. In *Vulnerability, Uncertainty, and Risk*, pages 895–904. 2014. doi: 10.1061/9780784413609.091.

[160] Tore Schweder and Nils Lid Hjort. Confidence and likelihood. *Scandinavian Journal of Statistics*, 29(2):309–332, 2002. doi: 10.1111/1467-9469.00285.

[161] Kesar Singh, Minge Xie, and William E. Strawderman. Combining information from independent sources through confidence distributions. *Annals of Statistics*, 33(1): 159–183, 2005. doi: 10.1214/009053604000001084.

[162] Ryan Martin. False confidence, non-additive beliefs, and valid statistical inference. *International Journal of Approximate Reasoning*, 113:39–73, 2019. doi: 10.1016/j.ijar.2019.06.005.

[163] Iain Carmichael and Jonathan Williams. An exposition of the false confidence theorem. *Stat*, 7(1):e201–e201, 2018. doi: 10.1002/sta4.201.

[164] Brian M. Adams, William J. Bohnhoff, Keith R. Dalbey, John P. Eddy, Micheal S. Eldred, David M. Gay, Karen Haskell, Patricia D. Hough, and Laura P. Swiler. DAKOTA, A Multilevel Parallel Object-Oriented Framework for Design Optimization, Parameter Estimation, Uncertainty Quantification, and Sensitivity Analysis: Version 5.0 Reference Manual. Technical report, Sandia National Laboratories, Albuquerque, NM, United States, 2010.

[165] Edoardo Patelli. COSSAN: A Multidisciplinary Software Suite for Uncertainty Quantification and Risk Management. In Roger Ghanem, David Higdon, and Houman Owhadi, editors, *Handbook of Uncertainty Quantification*, pages 1–69. Springer International Publishing, Cham, 2015. ISBN 978-3-319-11259-6. doi: 10.1007/978-3-319-11259-6_59-1.

[166] Stefano Marelli and Bruno Sudret. UQLab: A Framework for Uncertainty Quantification in Matlab. In *Vulnerability, Uncertainty, and Risk*, pages 2554–2563, Liverpool, UK, June 2014. American Society of Civil Engineers. ISBN 978-0-7844-1360-9. doi: 10.1061/9780784413609.257.

[167] Audrey Olivier, Dimitris G. Giovanis, B.S. Aakash, Mohit Chauhan, Lohit Vandanapu, and Michael D. Shields. UQpy: A general purpose Python package and

development environment for uncertainty quantification. *Journal of Computational Science*, 47:101204, November 2020. ISSN 18777503. doi: 10.1016/j.jocs.2020.101204.

[168] Scott Ferson. What Monte Carlo methods cannot do. *Human and Ecological Risk Assessment: An International Journal*, 2(4):990–1007, December 1996. ISSN 1080-7039, 1549-7860. doi: 10.1080/10807039609383659.

[169] Scott Ferson and Lev R Ginzburg. Different methods are needed to propagate ignorance and variability. *Reliability Engineering & System Safety*, 54(2–3):21, 1996.

[170] William Oberkampf. Simulation informed decision making [Conference Presentation]. In *Virtual Conference on Epistemic Uncertainty in Engineering*, 2021. https://www.youtube.com/watch?v=i4L3fUpr59s.

[171] Elisabeth Paté-Cornell. On ”Black Swans” and ”Perfect Storms”: Risk Analysis and Management When Statistics Are Not Enough. *Risk Analysis*, 32(11), 2012. doi: 10.1111/j.1539-6924.2011.01787.x.

[172] Richard P. Feynman. Appendix F - Personal Observations on Reliability of Shuttle. In *Report of the Presidential Commission on the Space Shuttle Challenger Accident*, volume 2. US Government Printing Office, Washington, DC, USA, 1986. https://history.nasa.gov/rogersrep/v2appf.htm.

[173] Yukiya Amano. The Fukushima Daiichi Accident Report by the Director General. Technical report, International Atomic Energy Agency, Vienna, Austria, 2015.

[174] Pravin K Trivedi and David M Zimmer. Copula Modeling: An Introduction for Practitioners. *Foundations and Trends in Econometrics*, 1(1):1–111, 2006. ISSN 1551-3076, 1551-3084. doi: 10.1561/0800000005.

[175] Ander Gray, Dominik Hose, Marco De Angelis, Michael Hanss, and Scott Ferson. Dependent Possibilistic Arithmetic using Copulas. In *Proceedings of the Twelth International Symposium on Imprecise Probabilities: Theories and Applications*, volume 147, pages 173–183, Granada, Spain (Virtual), 2021. Proceedings of Machine Learning Research.

[176] M D McKay, R J Beckman, and W J Conover. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics. American Statistical Association*, 21(2):239—245, 1979. doi: 10.2307/1268522.

[177] J.C. Helton and F.J. Davis. Latin hypercube sampling and the propagation of uncer-

tainty in analyses of complex systems. *Reliability Engineering & System Safety*, 81 (1):23–69, July 2003. ISSN 09518320. doi: 10.1016/S0951-8320(03)00058-9.

[178] Michael D. Shields and Jiaxin Zhang. The generalization of Latin hypercube sampling. *Reliability Engineering & System Safety*, 148:96–108, April 2016. ISSN 09518320. doi: 10.1016/j.ress.2015.12.002.

[179] Michael D. Shields, Kirubel Teferra, Adam Hapij, and Raymond P. Daddazio. Refined Stratified Sampling for efficient Monte Carlo based uncertainty quantification. *Reliability Engineering & System Safety*, 142:310–325, October 2015. ISSN 09518320. doi: 10.1016/j.ress.2015.05.023.

[180] J Jakeman, M Eldred, and D Xiu. Numerical approach for quantification of epistemic uncertainty. *Journal of Computational Physics*, 229(12):4648–4663, 2010. ISSN 0021-9991. doi: 10.1016/j.jcp.2010.03.003.

[181] Shuxing Yang, Fenfen Xiong, and Fenggang Wang. Polynomial Chaos Expansion for Probabilistic Uncertainty Propagation. In Jan Peter Hessling, editor, *Uncertainty Quantification and Model Calibration*. InTech, July 2017. ISBN 978-953-51-3279-0 978-953-51-3280-6. doi: 10.5772/intechopen.68484.

[182] Dirk P. Kroese, Tim Brereton, Thomas Taimre, and Zdravko I. Botev. Why the Monte Carlo method is so important today. *WIREs Computational Statistics*, 6(6): 386–392, November 2014. ISSN 1939-5108, 1939-0068. doi: 10.1002/wics.1314.

[183] Dailys Arronde Perez, Harald Gietler, and Hubert Zangl. Automatic Uncertainty Propagation Based on the Unscented Transform. In *2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, pages 1–6, Dubrovnik, Croatia, May 2020. IEEE. ISBN 978-1-72814-460-3. doi: 10.1109/I2MTC43012.2020.9129581.

[184] Paul Embrechts, Filip Lindskog, and Alexander Mcneil. Modelling Dependence with Copulas and Applications to Risk Management. In *Handbook of Heavy Tailed Distributions in Finance*, pages 329–384. Elsevier, 2003. ISBN 978-0-444-50896-6. doi: 10.1016/B978-044450896-6.50010-8.

[185] Roger B. Nelsen. *An Introduction to Copulas*. Springer Series in Statistics. Springer, New York, New York, USA, 2nd ed edition, 2006. ISBN 978-0-387-28659-4.

[186] Harry Joe. *Dependence Modeling with Copulas*. Chapman & Hall/CRC, Boca Raton, Florida, USA, 2014.

[187] Dominik Hose and Michael Hanss. A universal approach to imprecise probabilities

in possibility theory. *International Journal of Approximate Reasoning*, 133:133–158, June 2021. ISSN 0888613X. doi: 10.1016/j.ijar.2021.03.010.

[188] Terence Parr. *The Definitive ANTLR 4 Reference*. The Pragmatic Bookshelf, Dallas, USA, 2012. ISBN 978-1-934356-99-9.

[189] N S Nedialkov, K R Jackson, and G F Corliss. Validated solutions of initial value problems for ordinary differential equations. *Appl. Math. Comput.*, page 48, 1999.

[190] Nedialko S Nedialkov, Kenneth R Jackson, and John D. Pryce. An Effective High-Order Interval Method for Validating Existence and Uniqueness of the Solution of an IVP for an ODE. *Reliable Computing*, 7(6):17, 2001.

[191] Nedialko Nedialkov. Interval Tools for ODEs and DAEs. In *12th GAMM - IMACS International Symposium on Scientific Computing, Computer Arithmetic and Validated Numerics (SCAN 2006)*, pages 4–4, Duisburg, Germany, 2006. IEEE. ISBN 978-0-7695-2821-2. doi: 10.1109/SCAN.2006.28.

[192] Youdong Lin and Mark A Stadtherr. Validated solutions of initial value problems for parametric ODEs. *Applied Numerical Mathematics*, 57:1145–1162, 2007. doi: 10.1016/j.apnum.2006.10.006.

[193] Joshua A Enszer, D M Andrei, and Mark A Stadtherr. Probability bounds analysis for nonlinear population ecology models. *Mathematical Biosciences*, 267:97–108, 2015. doi: 10.1016/j.mbs.2015.06.012.

[194] David A. Plaisted. Source-to-Source Translation and Software Engineering. *Journal of Software Engineering and Applications*, 06(04):30–40, 2013. ISSN 1945-3116, 1945-3124. doi: 10.4236/jsea.2013.64A005.

[195] E Adams and U Kulisch. *Scientific Computing with Automatic Result Verification*. Academic Press, Boston, MA, USA, 1993. ISBN 978-0-12-410955-1.

[196] Gene G Hunder, Daniel A. Bloch, Beat A. Michel, Mary Betty Stevens, William P. Aren, Leonard H. Calabrese, Steven M. Edworthy, Anthony S. Fauci, Randi Y. Leavitt, J. T. Lie, Robert W. Lightfoot Jr., Alfonse T. Masi, Dennis J. McShane, John A. Mills, Stanley L. Wallace, and Nathan J. Zvaifler. The American College of Rheumatology 1990 criteria for the classification of giant cell arteritis. *Arthritis & Rheumatism*, 33(8):1122–1128, 1990.

[197] Luis von Ahn, Benjamin Maurer, Colin Mcmillen, David Abraham, and Manuel Blum. reCAPTCHA : Human-Based Recognition via Web Character. *Science*, 321(5895): 1465–1468, 2008.

[198] Alex Groce, Todd Kulesza, Chaoqiang Zhang, Shalini Shamasunder, Margaret Burnett, Weng-Keen Wong, Simone Stumpf, Shubhomoy Das, Amber Shinsel, Forrest Bice, and Kevin McIntosh. You Are the Only Possible Oracle: Effective Test Selection for End Users of Interactive Machine Learning Systems. *IEEE Transactions on Software Engineering*, 40(3):307–323, March 2014. ISSN 0098-5589, 1939-3520. doi: 10.1109/TSE.2013.59.

[199] Tzu-Kuo Huang, Lihong Li, Ara Vartanian, Saleema Amershi, and Jerry Zhu. Active learning with oracle epiphany. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, volume 29, Barcelona, Spain, 2016. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/299fb2142d7de959380f91c01c3a293c-Paper.pdf.

[200] P. Finbarr Duggan. Time to abolish 'gold standard'. *British Medical Journal*, 304 (6841):1568–1569, 1992.

[201] E. Versi. "Gold standard" is an appropriate term. *British Medical Journal*, 305(July): 1992–1992, 1992.

[202] John Collins and Paul S. Albert. Estimating diagnostic accuracy without a gold standard: A continued controversy. *Journal of Biopharmaceutical Statistics*, 26(6): 1078–1082, November 2016. ISSN 1054-3406, 1520-5711. doi: 10.1080/10543406.2016.1226334.

[203] David Coggon, Christopher Martyn, Keith T Palmer, and Bradley Evanoff. Assessing case definitions in the absence of a diagnostic gold standard. *International Journal of Epidemiology*, 34(4):949–952, August 2005. ISSN 1464-3685, 0300-5771. doi: 10.1093/ije/dyi012.

[204] Thomas Akkerhuis, Jeroen de Mast, and Tashi Erdmann. The statistical evaluation of binary tests without gold standard: Robustness of latent variable approaches. *Measurement*, 95:473–479, January 2017. ISSN 0263-2241. doi: 10.1016/j.measurement.2016.10.043.

[205] Anne Rutjes, J Reitsma, Arri Coomarasamy, Khalid Khan, and Patrick Bossuyt. Evaluation of diagnostic test when there is no gold standard. A review of methods. *Health Technology Assessment*, 11(50), 2007. doi: 10.3310/hta11500.

[206] Tanya Walsh. Fuzzy gold standards: Approaches to handling an imperfect reference standard. *Journal of Dentistry*, 74:S47–S49, July 2018. ISSN 0300-5712. doi: 10.1016/j.jdent.2018.04.022.

[207] Chinyereugo M. Umemneku Chikere, Kevin Wilson, Sara Graziadio, Luke Vale, and A. Joy Allen. Diagnostic test evaluation methodology: A systematic review of methods employed to evaluate diagnostic tests in the absence of gold standard – An update. *PLOS ONE*, 14(10):e0223832, October 2019. ISSN 1932-6203. doi: 10.1371/journal.pone.0223832.

[208] Chinyereugo M. Umemneku Chikere, Kevin J. Wilson, A. Joy Allen, and Luke Vale. Comparative diagnostic accuracy studies with an imperfect reference standard – a comparison of correction methods. *BMC Medical Research Methodology*, 21(1):67, December 2021. ISSN 1471-2288. doi: 10.1186/s12874-021-01255-4.

[209] Maurice Staquet, Marcel Rozencweig, Young Jack Lee, and Franco M. Muggia. Methodology for the assessment of new dichotomous diagnostic tests. *Journal of Chronic Diseases*, 34(12):599–610, January 1981. ISSN 00219681. doi: 10.1016/0021-9681(81)90059-X.

[210] John J Gart and Alfred A Buck. A Probabilistic Model for the Comparison of Diagnositc Tests. *American Journal of Epidemiology*, 85(3):10, 1965.

[211] Robert L. Winkler and James E. Smith. On Uncertainty in Medical Testing. *Medical Decision Making*, 24(6):654–658, 2004. ISSN 0272-989X. doi: 10.1177/0272989X04271045.

[212] John A. Baron. Uncertainty in Bayes. *Medical Decision Making*, 14(1):46–51, February 1994. doi: 10.1177/0272989X9401400106.

[213] E Lesaffre and AB Lawson. *Bayesian Biostatistics*. John Wiley & Sons, Ltd, Chichester, United Kingdom, 2012. doi.org/10.1002/9781119942412.

[214] Douglas Mossman and James O. Berger. Intervals for posttest probabilities: A comparison of 5 methods. *Medical Decision Making*, 21(6):498–507, 2001. ISSN 0102100608. doi: 10.1177/02729890122062857.

[215] James E. Smith and Robert L. Winkler. Casey's Problem: Interpreting and Evaluating a New Test. *Interfaces*, 29(3):63–76, June 1999. doi: 10.1287/inte.29.3.63.

[216] James E. Smith, Robert L. Winkler, and Dennis G. Fryback. The First Positive: Computing Positive Predictive Value at the Extremes. *Annals of Internal Medicine*, 132(10):804–804, May 2000. doi: 10.7326/0003-4819-132-10-200005160-00008.

[217] William M. Bolstad. *Introduction to Bayesian Statistics*. John Wiley & Sons, Ltd, Hoboken, N.J., USA, second edition, 2007. ISBN 978-0-470-18117-1.

[218] Peter Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, UK, 1991.

[219] Peter Walley. Inferences from Multinomial Data: Learning about a Bag of Marbles. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):3–57, 1996.

[220] Peter Walley, Lyle Gurrin, and Paul Burton. Analysis of Clinical Data Using Imprecise Prior Probabilities. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 45(4):457–485, 1996.

[221] Teddy Seidenfeld and Larry Wasserman. Dilation for Sets of Probabilities. *The Annals of Statistics*, 21(3):1139–1154, 1993.

[222] Sarah C Emerson, Sushrut S Waikar, Claudio Fuentes, Joseph V Bonventre, and Rebecca A Betensky. Biomarker validation with an imperfect reference: Issues and bounds. *Statistical Methods in Medical Research*, 27(10):2933–2945, October 2018. ISSN 0962-2802, 1477-0334. doi: 10.1177/0962280216689806.

[223] Hermann Brenner. Correcting for Exposure Misclassification Using an Alloyed Gold Standard:. *Epidemiology*, 7(4):406–410, July 1996. ISSN 1044-3983. doi: 10.1097/00001648-199607000-00011.

[224] Scott Menard. *Logistic Regression: From Introductory to Advanced Concepts and Applications*. SAGE Publications, Inc, Thousand Oaks, California, 2010. ISBN 978-1-4129-7483-7. doi: 10.4135/9781483348964.

[225] S James Press and Sandra Wilson. Choosing Between Logistic Regression and Discriminant Analysis. *Journal of the American Statistical Association*, 73(364):699–705, 1978.

[226] Steven C Bagley, Halbert White, and Beatrice A Golomb. Logistic regression in the medical literature : Standards for use and reporting , with particular attention to one medical domain. *Journal of Clinical Epidemiology*, 54:979–985, 2001.

[227] W D Neary, B P Heather, and J J Earnshaw. The Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity (POSSUM). *British Journal of Surgery*, 30(2):157–165, 2003. doi: 10.1002/bjs.4041.

[228] Sanjay Kumar Palei and Samir Kumar Das. Logistic regression model for prediction of roof fall risks in bord and pillar workings in coal mines : An approach. *Safety Science*, 47(1):88–96, 2009. doi: 10.1016/j.ssci.2008.01.002.

[229] Winifred Lambert and Mark Wheeler. Objective Lightning Probability Forecasting

for Kennedy Space Center and Cape Canaveral Air Force Station. Technical report, National Air and Space Administration, Hannover, MD, USA, 2005.

[230] Gregory C Ohlmacher and John C Davis. Using multiple logistic regression and GIS technology to predict landslide hazard in northeast Kansas, USA. *Engineering Geology*, 69:331–343, 2003. doi: 10.1016/S0013-7952(03)00069-3.

[231] Yongjun Li, Lizheng Wang, and Feng Li. A data-driven prediction approach for sports team performance and its application to National Basketball Association R. *Omega*, 98:102123–102123, 2021. doi: 10.1016/j.omega.2019.102123.

[232] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[233] Daniel Rabinowjtz, Anastasios Tsiatis, and Jorge Aragon. Regression with interval-censored data. *Biometrika*, 82(3):13, 1995.

[234] Jane C. Lindsey and Louise M. Ryan. Methods for interval-censored data. *Statistics in Medicine*, 17(2):219–238, January 1998. ISSN 0277-6715, 1097-0258. doi: 10.1002/(SICI)1097-0258(19980130)17:2<219::AID-SIM735>3.0.CO;2-O.

[235] L Billard and E Diday. From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis. *Journal of the American Statistical Association*, 98(462): 470–487, June 2003. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214503000242.

[236] T. Whitaker, B. Beranger, and S. A. Sisson. Logistic Regression Models for Aggregated Data. *Journal of Computational and Graphical Statistics*, 30(4):1049–1067, October 2021. ISSN 1061-8600, 1537-2715. doi: 10.1080/10618600.2021.1895816.

[237] Boris Beranger, Huan Lin, and Scott Sisson. New models for symbolic data analysis. *Advances in Data Analysis and Classification*, September 2022. ISSN 1862-5347, 1862-5355. doi: 10.1007/s11634-022-00520-8.

[238] Donald B Rubin. Inference and missing data. *Biometrika*, 63(8):12, 1976.

[239] Scott Ferson, Vladik Kreinovich, Janos Hajagos, William Oberkampf, and Lev Ginzburg. Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty. Technical report, Sandia National Laboratories, Albuquerque, NM, USA, 2007.

[240] Patrice Bertrand. Descriptive Statistics for Symbolic Data. In Hans-Hermann Bock and Edwin Diday, editors, *Analysis of Symbolic Data: Exploratory Methods for Ex-*

*tracting Statistical Information from Complex Data*, pages 106–124. Springer Berlin Heidelberg, 2000.

[241] L Billard and E Diday. Regression Analysis for Interval-Valued Data. In Henk A. L. Kiers, Jean-Paul Rasson, Patrick J. F. Groenen, and Martin Schader, editors, *Data Analysis, Classification, and Related Methods*, pages 369–374. Springer Berlin Heidelberg, 2000.

[242] HH Bock and E Diday. *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data, Edited by H.-H. Bock and E. Diday.* Springer Berlin Heidelberg, Berlin, Germany, 2000.

[243] JM Keynes. *A Treatise on Probability.* Macmillian and Co., London, UK, 1921.

[244] Renata M C R de Souza, Francisco Jos, A Cysneiros, Diego C F Queiroz, and Roberta A De A Fagundes. A Multi-Class Logistic Regression Model for Interval Data. In *IEEE International Conference on Systems, Man and Cybernetics*, pages 1253–1258, Singapore,Singapore, 2008. doi: 10.1109/ICSMC.2008.4811455.

[245] Renata M C R de Souza, Diego C F Queiroz, and Francisco Jose. Logistic regression-based pattern classifiers for symbolic interval data. *Pattern Analysis and Applications*, 14(3):273–282, 2011. doi: 10.1007/s10044-011-0222-1.

[246] CF Manski. *Partial Identification of Probability Distributions.* Springer, New York, NY USA, 2003. ISBN 0-387-00454-8.

[247] Hung T. Nguyen, Vladik Kreinovich, Berlin Wu, and Gang Xiang. *Computing Statistics under Interval and Fuzzy Uncertainty.* Springer, Heidelberg, Germany, 2012. ISBN 978-3-642-22829-2.

[248] Lev V Utkin and Frank P A Coolen. Interval-valued regression and classification models in the framework of machine learning. In *7th International Symposium on Imprecise Probability: Theories and Applications*, page 11, Innsbruck, Austria, 2011.

[249] Andrea Wiencierz. *Regression analysis with imprecise data.* PhD thesis, Ludwig-Maximilians-Universitat Munchen, 2013.

[250] Georg Schollmeyer. Computing Simple Bounds for Regression Estimates for Linear Regression with Interval-valued Covariates. In *Proceedings of the Twelth International Symposium on Imprecise Probabilities: Theories and Applications*, page 7, Virtual Conference, 2021. Proceedings of Machine Learning Research.

[251] Krasymyr Tretiak, Georg Schollmeyer, and Scott Ferson. Neural network model for

imprecise regression with interval dependent variables. http://arxiv.org/abs/2206.02467, June 2022.

[252] Roberta A A Fagundes, Renata M C R de Souza, and Francisco Jose. Robust regression with application to symbolic interval data. *Engineering Applications of Artificial Intelligence*, 26(1):564–573, 2013. doi: 10.1016/j.engappai.2012.05.004.

[253] Scott Ferson, Lev Ginzburg, Vladik Kreinovich, Luc Longpr, Monica Aviles, and North Country Road. Computing Variance for Interval Data is NP-Hard. *ACM SIGACT News*, 33(2):108–118, 2002.

[254] Massih-Reza Amini and Patrick Gallinari. Semi-Supervised Logistic Regression. In *15th European Conference on Artificial Intelligence*, page 5, Lyon, France, 2002.

[255] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, USA, ebook edition, 2006.

[256] Shengqiang Chi, Xinhang Li, Yu Tian, Jun Li, Xiangxing Kong, Kefeng Ding, Chunhua Weng, and Jingsong Li. Semi-supervised learning to improve generalizability of risk prediction models. *Journal of Biomedical Informatics*, 92:103117, April 2019. ISSN 15320464. doi: 10.1016/j.jbi.2019.103117.

[257] In Jae Myung. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47(1):90–100, 2003. doi: 10.1016/S0022-2496(02)00028-7.

[258] Tommi S Jaakkola. A variational approach to Bayesian logistic regression models and their extensions. In *Proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics*, page 12, Fort Lauderdale, FL, USA, 1997. PMLR.

[259] Sean M. O'Brien and David B. Dunson. Bayesian Multivariate Logistic Regression. *Biometrics*, 60(3):739–746, September 2004. ISSN 0006341X. doi: 10.1111/j.0006-341X.2004.00224.x.

[260] David G. Kleinbaum and Mitchel Klein. *Logistic Regression: A Self-Learning Text*. Springer, New York, NY USA, 3rd edition, 2010.

[261] Patrick Royston and Douglas G Altman. Visualizing and assessing discrimination in the logistic regression model. *Statistics in Medicine*, 29, 2010. doi: 10.1002/sim.3994.

[262] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009. doi: 10.1016/j.dss.2009.05.016.

[263] Balaji Krishnapuram, David Williams, Ya Xue, Lawrence Carin, Mário Figueiredo,

and Alexander J Hartemink. On Semi-Supervised Classification. In *Advances in Neural Information Processing Systems 17 (NIPS 2004)*, page 8, Vancouver/Whistler, Canada, 2004.

[264] Danilo Bzdok, Michael Eickenberg, Olivier Grisel, Bertrand Thirion, and Gael Varoquaux. Semi-Supervised Factored Logistic Regression for High-Dimensional Neuroimaging Data. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, page 9, Montreal, Canada, 2015.

[265] Hua Chai, Yong Liang, Sai Wang, and Hai-wei Shen. A novel logistic regression model combining semi-supervised learning and active learning for disease classification. *Scientific Reports*, 8(1):13009, December 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-31395-5.

[266] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics* -, pages 189–196, Cambridge, Massachusetts, 1995. Association for Computational Linguistics. doi: 10.3115/981658.981684.

[267] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Wine Quality Data Set. http://archive.ics.uci.edu/ml/datasets/Wine+Quality, 2009.

[268] Krasymyr Tretiak and Scott Ferson. Measuring uncertainty when pooling interval-censored data sets with different precision. http://arxiv.org/abs/2210.13863, October 2022.

[269] Andrew I. Schein and Lyle H. Ungar. Active learning for logistic regression: An evaluation. *Machine Learning*, 68(3):235–265, August 2007. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-007-5019-5.

[270] Department of Health and Social Care. COVID-19: Government announces moving out of contain phase and into delay phase. 2020. https://www.gov.uk/government/news/covid-19-government-announces-moving-out-of-contain-phase-and-into-delay.

[271] Department of Health and Social Care. Scaling up our testing programmes. Technical report, 2020. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/878121/coronavirus-covid-19-testing-strategy.pdf.

[272] Neil M Ferguson, Daniel Laydon, Gemma Nedjati-Gilani, Natsuko Imai, Kylie Ainslie, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, Zulma Cucunubá, Gina Cuomo-Dannenburg, Amy Dighe, Ilaria Dorigatti, Han Fu, Katy Gaythorpe, Will Green, Arran Hamlet, Wes Hinsley, Lucy C Okell, Sabine Van Elsland, Hay-

ley Thompson, Robert Verity, Erik Volz, Haowei Wang, Yuanrong Wang, Patrick Gt Walker, Caroline Walters, Peter Winskill, Charles Whittaker, Christl A Donnelly, Steven Riley, and Azra C Ghani. Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Technical report, 2020.

[273] Boris Johnson. PM address to the nation on coronavirus: 23 March 2020. 2020. https://www.gov.uk/government/speeches/pm-address-to-the-nation-on-coronavirus-23-march-2020.

[274] Alasdair Sandford. Coronavirus: Half of humanity now on lockdown as 90 countries call for confinement. *Euronews*, April 2020. https://www.euronews.com/2020/04/02/coronavirus-in-europe-spain-s-death-toll-hits-10-000-after-record-950-new-deaths-in-24-hou.

[275] UNCESCO. COVID-19 Impact on Eduction. https://en.unesco.org/covid19/educationresponse, 2020.

[276] Kiesha Prem, Yang Liu, Timothy W Russell, Adam J Kucharski, Rosalind M Eggo, Nicholas Davies, Mark Jit, Petra Klepac, Stefan Flasche, Samuel Clifford, Carl A B Pearson, James D Munday, Sam Abbott, Hamish Gibbs, Alicia Rosello, Billy J Quilty, Thibaut Jombart, Fiona Sun, Charlie Diamond, Amy Gimma, Kevin van Zandvoort, Sebastian Funk, Christopher I Jarvis, W John Edmunds, Nikos I Bosse, and Joel Hellewell. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: A modelling study. *The Lancet Public Health*, 2667(20):1–10, 2020. doi: 10.1016/s2468-2667(20)30073-6.

[277] Kathy Leung, Joseph T Wu, Di Liu, and Gabriel M Leung. First-wave COVID-19 transmissibility and severity in China outside Hubei after control measures, and second-wave scenario planning: A modelling impact assessment. *The Lancet*, 6736 (20), 2020. doi: 10.1016/S0140-6736(20)30746-7.

[278] Harriet Agerholm and Dulcie Lee. Boris Johnson: Do not throw caution to the wind on Covid. *BBC News*, February 2022. https://www.bbc.co.uk/news/uk-60446908.

[279] Richard Horton. Offline: COVID-19—bewilderment and candour. *The Lancet*, 395 (10231):1178–1178, 2020. doi: 10.1016/S0140-6736(20)30850-3.

[280] Richard Horton. Offline : COVID-19 and the NHS —“a national scandal”. *The Lancet*, 395(10229):1022–1022, 2020. doi: 10.1016/S0140-6736(20)30727-3.

[281] Tedros Adhanom. WHO Director-General’s opening remarks at the media briefing on COVID-19 - 16 March 2020. https://www.who.int/dg/speeches/detail/

who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---16-march-2020, 2020.

[282] Matt Hancock and Department of Health and Social Care. Press release - Health Secretary sets out plan to carry out 100,000 coronavirus tests a day. https://www.gov.uk/government/news/health-secretary-sets-out-plan-to-carry-out-100000-coronavirus-tests-a-day, April 2020.

[283] 10 Downing Street. Coronavirus press conference (2 April 2020). https://www.youtube.com/watch?v=EY4Hr5-fk9c, 2020.

[284] Eleanor Cummins. Why the Coronavirus Test Gives So Many False Negatives. https://slate.com/technology/2020/04/coronavirus-testing-false-negatives.html, 2020.

[285] Qinjian Hao, Hongmei Wu, and Qiang Wang. Difficulties in False Negative Diagnosis of Coronavirus Disease 2019: A Case Report. *Infectious Diseases - Preprint*, pages 1–12, 2020. doi: 10.21203/RS.3.RS-17319/V1.

[286] Harlan M. Krumholz. If You Have Coronavirus Symptoms, Assume You Have the Illness, Even if You Test Negative. *New York Times*, April 2020. https://www.nytimes.com/2020/04/01/well/live/coronavirus-symptoms-tests-false-negative.html.

[287] Karina Lichtenstein. Are Coronavirus Tests Accurate? *MedicineNet Health News*, 2020. https://www.medicinenet.com/script/main/art.asp?articlekey=228250.

[288] Cormac Sheridan. Coronavirus and the race to distribute reliable diagnostics. *Nature Biotechnology*, 38(April):379–391, 2020. doi: 10.1038/d41587-020-00003-1.

[289] Giorgia Guglielmi. Fast Coronavirus Tests Are Coming. *Nature News Feature*, 2020. https://www.nature.com/articles/d41586-020-02661-2.

[290] L.E. Smith, H.W.W. Potts, R. Amlot, N.T. Fear, S. Michie, and G.J. Rubin. Do members of the public think they should use lateral flow tests (LFT) or polymerase chain reaction (PCR) tests when they have COVID-19-like symptoms? The COVID-19 Rapid Survey of Adherence to Interventions and Responses study. *Public Health*, 198:260–262, September 2021. ISSN 00333506. doi: 10.1016/j.puhe.2021.07.023.

[291] Anna Petherick. Developing antibody tests for SARS-CoV-2. *The Lancet*, 395(10230): 1101–1102, 2020. doi: 10.1016/S0140-6736(20)30788-1.

[292] Ben Riley-Smith and Sarah Knapton. Boris Johnson and Donald Trump talk up potential 'game-changer' scientific advances on coronavirus. *The*

*Telegraph*, March 2020. https://www.telegraph.co.uk/news/2020/03/20/boris-johnson-donald-trump-talk-potential-game-changer-scientific/.

[293] Nicholas C Grassly, Marga Pons-salort, Edward P K Parker, Peter J White, Kylie Ainslie, and Marc Baguelin. Report 16 : Role of testing in COVID-19 control. (April):1–13, 2020.

[294] Mark A. Hall and David M. Studdert. “Vaccine Passport” Certification — Policy and Ethical Considerations. *New England Journal of Medicine*, 385(11):e32, September 2021. ISSN 0028-4793, 1533-4406. doi: 10.1056/NEJMp2104289.

[295] Simon Hodes and Azeem Majeed. Using the NHS App as a covid-19 vaccine passport. *BMJ*, page n1178, May 2021. ISSN 1756-1833. doi: 10.1136/bmj.n1178.

[296] Androula Pavli and Helena C Maltezou. COVID-19 vaccine passport for safe resumption of travel. *Journal of Travel Medicine*, 28(4):taab079, June 2021. ISSN 1195-1982, 1708-8305. doi: 10.1093/jtm/taab079.

[297] D B Jernigan, S L Lindstrom, J R Johnson, J D Miller, M Hoelscher, R Humes, R Shively, L Brammer, S A Burke, J M Villanueva, A Balish, T Uyeki, D Mustaquim, A Bishop, J H Handsfield, R Astles, X Xu, A I Klimov, N J Cox, and M W Shaw. Detecting 2009 Pandemic Influenza A ( H1N1 ) Virus Infection : Availability of Diagnostic Testing Led to Rapid Pandemic Response. *Clinical Infectious Diseases*, 52(Suppl 1):S36–S43, 2011. doi: 10.1093/cid/ciq020.

[298] Bi Qifang, Wu Yongsheng, Mei Shujiang, Ye Chenfei, Zou Xuan, Zhang Zhen, Liu Xiaojian, Wei Lan, A Truelove Shaun, Zhang Tong, Gao Wei, Cheng Cong, Tang Xiujuan, Wu Xiaoliang, Wu Yu, Sun Binbin, Huang Suli, Sun Yu, Zhang Juncen, Ma Ting, Lessler Justin, and Feng Tiejian. Epidemiology and Transmission of COVID-19 in Shenzhen China: Analysis of 391 cases and 1,286 of their close contacts. *medRxiv Pre-print*, 3099(20):1–9, 2020. doi: 10.1016/S1473-3099(20)30287-5.

[299] Qun Li, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, Kathy S.M. Leung, Eric H.Y. Lau, Jessica Y. Wong, Xuesen Xing, Nijuan Xiang, Yang Wu, Chao Li, Qi Chen, Dan Li, Tian Liu, Jing Zhao, Man Liu, Wenxiao Tu, Chuding Chen, Lianmei Jin, Rui Yang, Qi Wang, Suhua Zhou, Rui Wang, Hui Liu, Yinbo Luo, Yuan Liu, Ge Shao, Huan Li, Zhongfa Tao, Yang Yang, Zhiqiang Deng, Boxi Liu, Zhitao Ma, Yanping Zhang, Guoqing Shi, Tommy T.Y. Lam, Joseph T. Wu, George F. Gao, Benjamin J. Cowling, Bo Yang, Gabriel M. Leung, and Zijian Feng. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected

Pneumonia. *New England Journal of Medicine*, pages 1199–1207, 2020. doi: 10.1056/nejmoa2001316.

[300] Daniel F Gudbjartsson, Agnar Helgason, Hakon Jonsson, Olafur T Magnusson, Pall Melsted, Gudmundur L Norddahl, Jona Saemundsdottir, Asgeir Sigurdsson, Patrick Sulem, Arna B Agustsdottir, Berglind Eiriksdottir, Run Fridriksdottir, Elisabet E Gardarsdottir, Gudmundur Georgsson, Olafia S Gretarsdottir, Kjartan R Gudmundsson, Thora R Gunnarsdottir, Arnaldur Gylfason, Hilma Holm, Brynjar O Jensson, Aslaug Jonasdottir, Frosti Jonsson, Kamilla S Josefsdottir, Thordur Kristjansson, Droplaug N Magnusdottir, Louise le Roux, Gudrun Sigmundsdottir, Gardar Sveinbjornsson, Kristin E Sveinsdottir, Maney Sveinsdottir, Emil A Thorarensen, Bjarni Thorbjornsson, Arthur Löve, Gisli Masson, Ingileif Jonsdottir, Alma D Möller, Thorolfur Gudnason, Karl G Kristinsson, Unnur Thorsteinsdottir, and Kari Stefansson. Spread of SARS-CoV-2 in the Icelandic Population. *New England Journal of Medicine*, pages 1–14, 2020. doi: 10.1056/nejmoa2006100.

[301] Giulia Giordano, Franco Blanchini, Raffaele Bruno, Patrizio Colaneri, Alessandro Di Filippo, Angela Di Matteo, and Marta Colaneri. *Modelling the COVID-19 Epidemic and Implementation of Population-Wide Interventions in Italy.* 2020. ISBN 4159102008837. doi: 10.1038/s41591-020-0883-7.

[302] Stephen M Kissler, Christine Tedijanto, Edward Goldstein, Yonatan H Grad, and Marc Lipsitch. Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period. *Science*, 2020. doi: 10.1126/science.abb5793.

[303] Korea Center for Disease Control and Prevention. Updates on COVID-19 in Republic of Korea. Technical report, Soeul, South Korea, 2020. https://www.cdc.go.kr/board/board.es?mid=a30402000000&bid=0030&act=view&list_no=366817&tag=&nPage=1.

[304] Zhengtu Li, Yongxiang Yi, Xiaomei Luo, Nian Xiong, Yang Liu, Shaoqiang Li, Ruilin Sun, Yanqun Wang, Bicheng Hu, Wei Chen, Yongchen Zhang, Jing Wang, Baofu Huang, Ye Lin, Jiasheng Yang, Wensheng Cai, Xuefeng Wang, Jing Cheng, Zhiqiang Chen, Kangjun Sun, Weimin Pan, Zhifei Zhan, Liyan Chen, and Feng Ye. Development and Clinical Application of A Rapid IgM-IgG Combined Antibody Test for SARS-CoV-2 Infection Diagnosis. *Journal of Medical Virology*, (February), 2020. doi: 10.1002/jmv.25727.

[305] Yang Yang, Minghui Yang, Chenguang Shen, Fuxiang Wang, and Jing Yuan. Evaluating the accuracy of different respiratory specimens in the laboratory diagnosis

and monitoring the viral shedding of 2019-nCoV infections ABSTRACT :. *MedRxiv Pre-Print*, 2020. doi: 10.1101/2020.02.11.20021493.

[306] Division of Viral Diseases National Center for Immunization and Respiratory Diseases (NCIRD). Interim Guidelines for Collecting, Handling, and Testing Clinical Specimens from Persons for Coronavirus Disease 2019 (COVID-19). 2020. https://www.cdc.gov/coronavirus/2019-ncov/lab/guidelines-clinical-specimens.html.

[307] Tao Ai, Zhenlu Yang, and Liming Xia. Correlation of Chest CT and RT-PCR Testing in Coronavirus Disease. *Radiology*, 2019:1–8, 2020. doi: 10.14358/PERS.80.2.000.

[308] Lorena Porte, Paulette Legarraga, Valeska Vollrath, Ximena Aguilera, José M Munita, Rafael Araos, Gabriel Pizarro, Pablo Vial, Sabine Dittrich, and Thomas Weitzel. Evaluation of Novel Antigen-Based Rapid Detection Test for the Diagnosis of SARS-CoV-2 in Respiratory Samples. *Available at SSRN 3569871*, 2020.

[309] Yang Pan, Luyao Long, Daitao Zhang, Tingting Yan, Shujuan Cui, Peng Yang, Quanyi Wang, and Simei Ren. Potential false-negative nucleic acid testing results for Severe Acute Respiratory Syndrome Coronavirus 2 from thermal inactivation of samples with low viral loads. *Clinical Chemistry*, 2020. doi: 10.1093/clinchem/hvaa091.

[310] Irene Cassaniti, Federica Novazzi, Federica Giardina, Francesco Salinaro, Michele Sachs, Stefano Perlini, Raffaele Bruno, Francesco Mojoli, Fausto Baldanti, and Members of the San Matteo Pavia COVID-19 Task Force. Performance of VivaDiag COVID - 19 IgM/IgG Rapid Test is inadequate for diagnosis of COVID-19 in acute patients referring to emergency room department. *Journal of Medical Virology*, (April):1–4, 2020. doi: 10.1002/jmv.25800.

[311] M Döhla, C Boesecke, B Schulte, C Diegmann, E Sib, E Richter, M Eschbach-Bludau, S Aldabbagh, B Marx, A M Eis-Hübinger, R M Schmithausen, and H Streeck. Rapid point-of-care testing for SARS-CoV-2 in a community screening setting shows low sensitivity. *Public Health*, 182:170–172, 2020. doi: 10.1016/j.puhe.2020.04.009.

[312] Laure Wynants, Ben Van Calster, Marc M J Bonten, Gary S Collins, Thomas P A Debray, Maarten De Vos, Maria C Haller, Georg Heinze, Karel G M Moons, Richard D Riley, Ewoud Schuit, Luc J M Smits, Kym I E Snell, Ewout W Steyerberg, Christine Wallisch, and Maarten Van Smeden. Prediction models for diagnosis and prognosis of covid-19 infection: Systematic review and critical appraisal. *The BMJ*, 369, 2020. doi: 10.1136/bmj.m1328.

[313] FIND. SARS-COV-2 DIAGNOSTICS: PERFORMANCE DATA. https://www.finddx.org/covid-19/dx-data/, 2020.

[314] PHE Porton Down and University of Oxford. Preliminary report from the Joint PHE Porton Down & University of Oxford SARS-CoV-2 test development and validation cell: Rapid evaluation of Lateral Flow Viral Antigen detection devices (LFDs) for mass community testing. Technical report, Wiltshire, United Kingdom, 2020.

[315] Jonathan J Deeks and Angela E Raffle. Lateral flow tests cannot rule out SARS-CoV-2 infection. *BMJ*, page m4787, December 2020. ISSN 1756-1833. doi: 10.1136/bmj.m4787.

[316] Jon Deeks, Angela Raffle, and Mike Gill. Covid-19: Government must urgently rethink lateral flow test roll out. *BMJ*, January 2021. https://blogs.bmj.com/bmj/2021/01/12/covid-19-government-must-urgently-rethink-lateral-flow-test-roll-out/.

[317] Jacqui Wise. Covid-19: Lateral flow tests miss over half of cases, Liverpool pilot data show. *BMJ*, page m4848, December 2020. ISSN 1756-1833. doi: 10.1136/bmj.m4848.

[318] UK Health Security Agency. Testing in England | Coronavirus in the UK. https://coronavirus.data.gov.uk/details/testing?areaType=nation&areaName=England, 2022.

[319] Herbert W Hethcote. The Mathematics of Infectious Diseases. *SIAM Review*, 42(4): 599–653, 2000.

[320] Sahar Sotoodeh Ghorbani, Niloufar Taherpour, Sahar Bayat, Hadis Ghajari, Parisa Mohseni, and Seyed Saeed Hashemi Nazari. Epidemiologic characteristics of cases with reinfection, recurrence, and hospital readmission due to COVID-19: A systematic review and meta-analysis. *Journal of Medical Virology*, 94(1):44–53, January 2022. ISSN 0146-6615, 1096-9071. doi: 10.1002/jmv.27281.

[321] Louise E Smith, Henry W W Potts, Richard Amlot, Nicola T Fear, Susan Michie, and G James Rubin. Adherence to the test, trace, and isolate system in the UK: Results from 37 nationally representative surveys. *BMJ*, page n608, March 2021. ISSN 1756-1833. doi: 10.1136/bmj.n608.

[322] Neel V. Patel. Why it's too early to start giving out "immunity passports". https://www.technologyreview.com/2020/04/09/998974/immunity-passports-cornavirus-antibody-test-outside/, 2020.

[323] Kate Proctor, Ian Sample, and Philip Oltermann. 'Immunity passports' could speed up return to work after Covid-19. https://www.theguardian.com/world/2020/mar/30/immunity-passports-could-speed-up-return-to-work-after-covid-19, 2020.

[324] The Times. Britain has millions of coronavirus antibody tests,

but they don't work. 2020. https://www.thetimes.co.uk/article/britain-has-millions-of-coronavirus-antibody-tests-but-they-don-t-work-j7kb55g89.

[325] Ying Liu, Albert A Gayle, Annelies Wilder-Smith, and Joacim Rocklöv. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine*, (Figure 1):1–4, 2020. doi: 10.1093/jtm/taaa021.

[326] Julien Riou and Christian L. Althaus. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin*, 25(4):1–5, 2020. doi: 10.2807/1560-7917.ES.2020.25.4.2000058.