Unravelling the interaction between somatic genetics and treatment response in leukaemia using machine learning

Ellie Rose Butler

A thesis submitted in part requirement for the degree of Doctor of Philosophy from the Faculty of Medical Sciences

Leukaemia Research Cytogenetics Group

Wolfson Childhood Cancer Research Centre

Newcastle University Centre for Cancer

Translational and Clinical Research Institute

Faculty of Medical Science

Newcastle University

September 2024

Declaration

The material documented in this thesis has not been submitted for a degree or qualification in this or any other university. I declare that the work carried out in this thesis is my own unaided work except where it is acknowledged otherwise.

Ellie Rose Butler

September 2024

Acknowledgements

Firstly, I would like to thank my supervisors Professor Anthony Moorman and Dr Amir Enshaei for their guidance and support throughout my PhD. I would also like to thank the past and present members of the Leukaemia Research Cytogenetics Group for their input, advice, and impartation of scientific knowledge during my time as both a research assistant and a PhD student. I would also like to thank them for their hard work producing and collating much of the data analysed in this thesis. Furthermore, I would like to thank my friends in the Wolfson Childhood Cancer Research Centre for their advice and reassurances; as well as the regular tea breaks and pub trips to keep me sane – especially She-Ra's Sidekicks: Zoe, Ruth, Jack, Stacey, and Dean. I would like to extend my thanks to my panel members Professor Christine Harrison and Dr Phillip Lord for their valued insight and advice during my PhD.

Next I would like to express my gratitude to my family for their encouragement and interest. In particular I would like to thank my mum and dad for their unwavering support, encouragement, and advice in the form of many phone calls on my walks home. You have always quashed my self-doubts and assured me I am good enough; whilst giving me every opportunity and supporting me in every venture. I would not be where I am, nor the person I am, today without you both and for that I will forever be immensely grateful. To Luke, thank you for being my constant support and providing never-ending comfort and reassurance. To Oreo, you are my favourite thing in this world. Seeing your little face and watching your zoomies when I get home after a long day makes everything better.

Additionally, I would like to thank my funder, the Alice Stephenson Bequest fund for making this project possible.

Lastly, I would like to dedicate this work to my Nannan, Christine Wendy Butler. You instilled in me a love of maths and inspired me to help others suffering from cancer. May you forever rest in peace.

Abstract

Acute lymphoblastic leukaemia (ALL) is the most common type of cancer affecting children with a peak prevalence between the ages of 2 and 5 years old. Development of effective treatments and improvements in risk stratification has led to a cure rate >90% in children. However, there are significant long-term effects associated with treatment of ALL. As a result, current research efforts have focused increasingly on identifying patients eligible for treatment de-escalation. Recent studies suggest that modest de-escalation of treatment for low risk patients is safe, namely patients with low levels of MRD and good risk genetics (*ETV6*::*RUNX1* and High Hyperdiploidy).

The objectives of this project were to identify optimal treatment elements for patients with good risk genetics to ensure that these patients are given only the minimal dosages of drugs necessary to be cured.

The survival rates of good risk genetics (*ETV6*::*RUNX1* and high hyperdiploidy) patients across four UKALL trials were determined, and the impact of different treatment elements was assessed individually using traditional statistical techniques to identify optimal treatment elements for these patients. Individual drug dosages were calculated using the trial protocols and a clinically annotated dataset (n = 6716) was assembled from both this information and data from LRCG sources. Area under the curve (AUC) was used to produce a dose intensity score (DIS) which was utilised as a method for determining optimal drug dosages for patients. Machine learning methods were explored with classification decision tree models being produced as well as ensemble methods being employed to identify optimal treatment elements within the aforementioned trials and analyse treatment effect on survival.

In conclusion, successful treatment pathways that optimise outcome and minimise toxicity exist within historic clinical trials. Furthermore, optimal doses of drugs given on current treatment protocols have been identified for good risk ALL patients.

Table of Contents

List of figures

List of Tables

Abbreviations

Abbreviation	Definition
ADTree	Alternating decision tree
AI	Artificial intelligence
AIEOP-BFM	Associazione Italiana di Ematologia e Oncologia Pediatrica and Berlin Frankfurt Münster
ALL	Acute lymphoblastic leukaemia
ALL IC-BFM	Acute Lymphoblastic Leukemia Intercontinental-Berlin Frankfurt Münster
Allo-HSCT	Allogeneic haematopoietic stem cell transplantation
Allo-SCT	Allogenic stem cell transplant
ANC	Absolute neutrophil count
AUC	Area under the curve
BFM	Berlin Frankfurt Münster
CCG	Children's Cancer Group
CCLSG	Children's cancer and Leukemia Study Group
CI	Confidence interval
CIMS	Cytogenetic information management system
CNA	Copy number alterations
CNN	Convolutional neural network
CNS	Central nervous system
CoALL	Childhood Acute Lymphoblastic Leukemia
CR	Complete remission

CRT	Cranial radiotherapy
DCNN	Deep convolutional neural network
DCOG	Dutch Childhood Oncology Group
DI	Delayed intensification
DNA	Deoxyribonucleic acid
DT	Decision Tree
DT	Double trisomy
EFS	Event-free survival
EOI	End of induction
FISH	Fluorescence in-situ hybridisation
FNR	False negative rate
FPR	False positive rate
Gini	Generalised inequality index
Gy	Gray
HCA	Hierarchical clustering analysis
HDM	High dose methotrexate
HD-MTX	High dose methotrexate
HeH	High hyperdiploidy
HR	Hazard ratio
iAMP21	Intrachromosomal Amplification of chromosome 21
ICCC-3	International classification of childhood cancer, third edition
IT	Intrathecal
Iu	International units

JACLS	Japan Childhood Leukemia Study Group
KNN	Kth-nearest neighbour
LOO	Leave-one-out
LRCG	Leukaemia research cytogenetics group
m	Metres
mg	Milligrams
ML	Machine learning
MLPA	Multiplex ligation-dependent probe amplification
MRC	Medial Research Council
MRD	Measurable residual disease
MRI	Magnetic resonance imaging
MTX	Methotrexate
NC	Not classified
NCI	National Cancer Institute
NE	Not eligible
NGS	Next-generation sequencing
NHS	National health service
NOPHO	Nordic Society of Pediatric Hematology and Oncology
OOB	Out of bag
OS	Overall survival
PCR	Polymerase chain reaction
PCR	Polymerase chain reaction
POG	Pediatric Oncology Group

rDRI	Refined disease risk index
Rel/ref	Relapse/refractory disease
RF	Random Forest
RNA	Ribonucleic acid
ROC	Receiving operating characteristic
RT-PCR	Real time- Polymerase chain reaction
SER	Slow early response
SMOTE	Synthetic minority over-sampling technique
SNP	Single nucleotide polymorphism
SVM	Support vector machine
TKI	Tyrosine kinase inhibitor
TNR	True negative rate
TPR	True positive rate
t-SNE	T-stochastic Neighbour Embedding
TYA	Teenagers and young adults
UKALL	United Kingdom Acute Lymphoblastic Leukaemia
VOD	Veno-Occlusive disease
WBC	White blood cell
WHO	World health organisation
XGBoost	eXtreme gradient boosting
µl	microlitre

Chapter 1. Introduction

1.1 Paediatric Cancer

Paediatric cancer is a rare event that accounts for less than 1% of all diagnosed cancer cases with an average of ~1800 new cases being diagnosed each year in the UK. Almost half (45%) of these cases occur in patients under the age of 5. Incidence rates have increased over time with a 12% positive trend since the early 1990s (Cancer Research UK, 2024b). Due to multimodal therapy survival rates for patients have greatly improved, with rates increasing from 36% in the 1970s to a 5-year survival rate of 84% in 2016 (Figure 1). However, survival differs by cancer with poor prognosis still present in certain tumour types. Although paediatric cancer is rare with largely positive outcomes, it still accounts for almost a quarter (23%) of deaths in children aged 1-14 (Erdmann *et al.*, 2021; Cancer Research UK, 2024b).

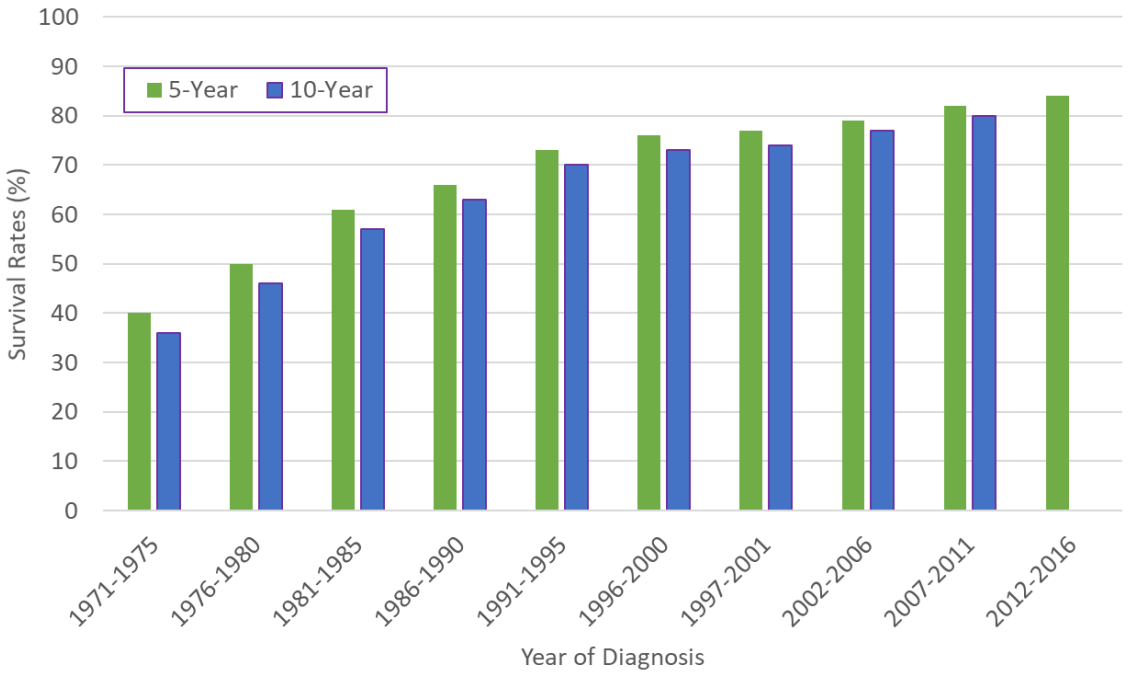


Figure 1. 5-year and 10-year survival rates (%) of paediatric cancer patients from 1971-2000 (Great Britain) and 1997-2016 (United Kingdom) amongst children aged 0-14 years old. Data from: (Cancer Research UK, 2024b).

The international classification of childhood cancer, third edition (ICCC-3) have classified childhood cancers into 12 diagnostic groups (Steliarova-Foucher *et al.*, 2005). The most

common types of paediatric cancer in the UK are leukaemias (31%), brain and spinal tumours (25%) and lymphomas (10%). The incidences of the 12 groups are shown in Figure 2.

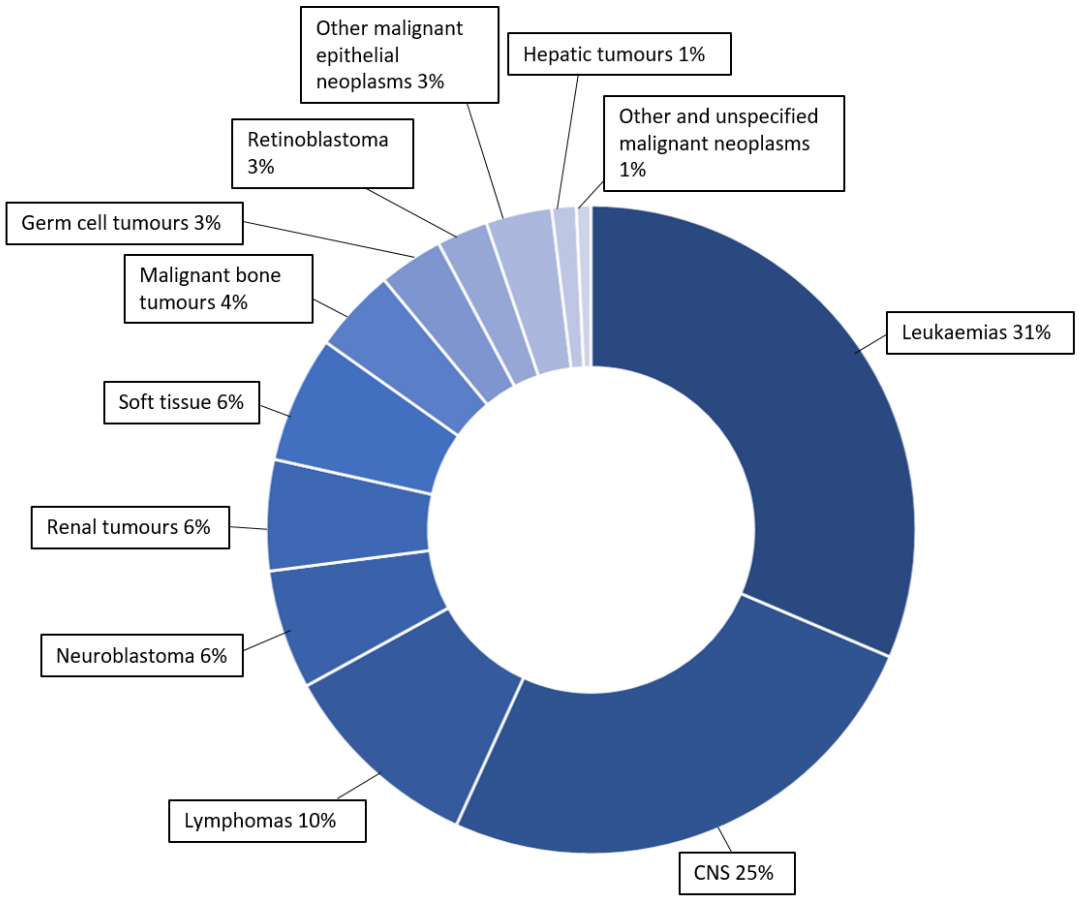


Figure 2. Incidence (%) of 12 paediatric cancer diagnostic groups classified by the ICCC-3 in children aged 0-14 years old in the United Kingdom (1997-2016). Data from: (Public Health England, 2021).

Although improvements in survival rates of paediatric cancer patients is a great positive, it has led to a growing population of survivors whom, due to intensive treatment at a young age, are now at risk of serious somatic and mental health conditions and ultimately, a reduced overall quality of life. Cancer treatment including chemotherapy, radiation, and stem cell transplantation can exert deleterious effects on normal human function and consequently increase the risk for early mortality, cardiac impairments, sensory loss, gastrointestinal problems, neurocognitive deficits, musculoskeletal abnormalities, and infertility (Ness and

Gurney, 2007; Erdmann *et al.*, 2021; Bhakta *et al.*, 2017; Butler and Haser, 2006; Friedman and Meadows, 2002). These sequelae highlight the need for alternative treatment approaches and improved risk stratification for patients on modern treatment protocols.

1.2 Leukaemia

Leukaemia is a type of cancer caused by the unregulated proliferation of a clone of immature blood cells derived from mutant haematopoietic stem cells (Howard and Hamilton, 2013). Leukaemia can be differentiated into acute or chronic depending on the degree of haematopoietic cell differentiation as well as the speed at which the disease progresses. It can be classified by cell lineage resulting in myeloid or lymphoid leukaemias (Loke and Kansagra, 2022). This defines four main types of leukaemia: acute myeloid leukaemia, acute lymphoblastic leukaemia, chronic myeloid leukaemia, and chronic lymphocytic leukaemia. Only acute lymphoblastic leukaemia will be considered in this thesis.

1.3 Acute Lymphoblastic Leukaemia

1.3.1 Diagnosis

Acute Lymphoblastic Leukaemia (ALL) is caused by an overproduction of immature lymphoid cells affecting both adults and children with a peak prevalence between 2-5 years old (Pui, Robison and Look, 2008). Figure 3 illustrates the normal haematopoiesis from a stem cell to a mature cell with the lymphoid arm depicted on the right hand side. The stage of differentiation at which overproduction occurs is highlighted. Diagnosis of ALL usually occurs due to clinical suspicion based on features and symptoms rather than incidentally, such as a result of a blood test (Bain, 2017). Symptoms of the disease occur from ALL cells infiltrating tissues or from varying degrees of anaemia, neutropenia, and thrombocytopenia (Kebriaei, Anastasi and Larson, 2002). The most common sites for ALL to be detected clinically are the lymph nodes, central nervous system (CNS), spleen, liver, and skin; though most organ systems may be involved after the entrance of leukaemia cells to the peripheral blood (Kebriaei, Anastasi and Larson, 2002). Presenting features of ALL include pallor, fever or other signs of infection, pharyngitis, petechiae, bone pain, hepatomegaly, splenomegaly, lymphadenopathy, gum hypertrophy, bruising and bleeding easily, shortness of breath, unexplained weight loss, and skin infiltration (Bain, 2017; Cancer Research UK, 2024b; Kebriaei, Anastasi and Larson, 2002). Diagnosis of ALL is based on WHO classification

guidelines which integrate the characterisations of cell morphology, immunophenotypes and genetics (Malard and Mohty, 2020).

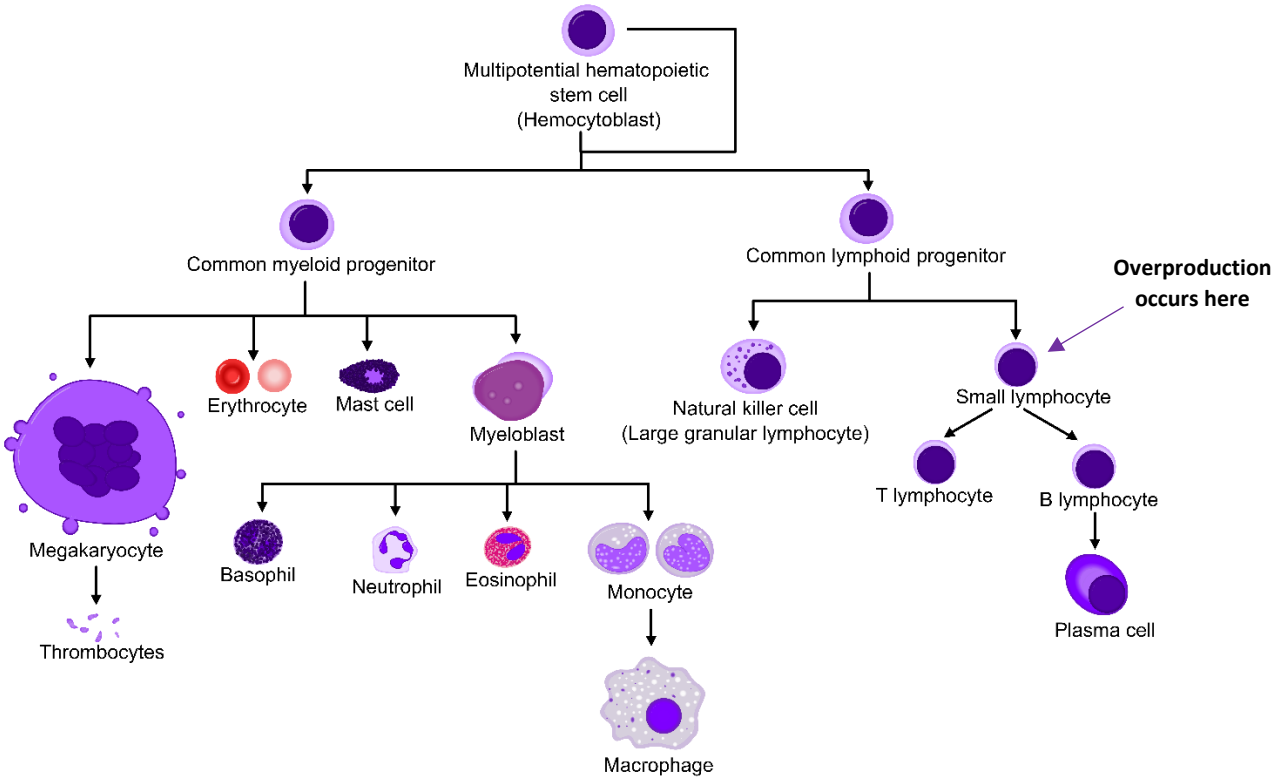


Figure 3. Diagram of the development of different blood cells from a haematopoietic stem cell to mature cells. (Rad and Häggström, 2009).

1.3.2 Etiology

The cause of acute leukaemias remains largely unknown (Tebbi, 2021). However, it is thought to arise from interactions between intrinsic and extrinsic exposures, genetic susceptibility, and chance as exemplified in Figure 4 (Inaba, Greaves and Mullighan, 2013). A commonly accepted hypothesis is the "multi-hit hypothesis" which states that an initial genetic hit must be followed by additional hits before a cell becomes leukaemic (Lausten-Thomsen *et al.*, 2010). As fusion genes and aneuploidy have been detected in the neonatal blood spots of children diagnosed with ALL, and *ETV6*::*RUNX1* B-ALL has developed in monozygotic twins, many believe the primary hit is a pre-leukaemic clone carrying a genetic lesion that was acquired *in utero* (Swaminathan *et al.*, 2015; Williams *et al.*, 2019; Alpar *et al.*, 2015). *In utero* development has been shown for several B-cell precursor ALL subtypes including high

hyperdiploidy, *ETV6*::*RUNX1*, *BCR*::*ABL1*, and *KMT2A* rearrangements (Rüchel *et al.*, 2022). During early childhood, pre-leukaemic clones can acquire secondary somatic mutations and evolve towards overt leukaemia (Swaminathan *et al.*, 2015).

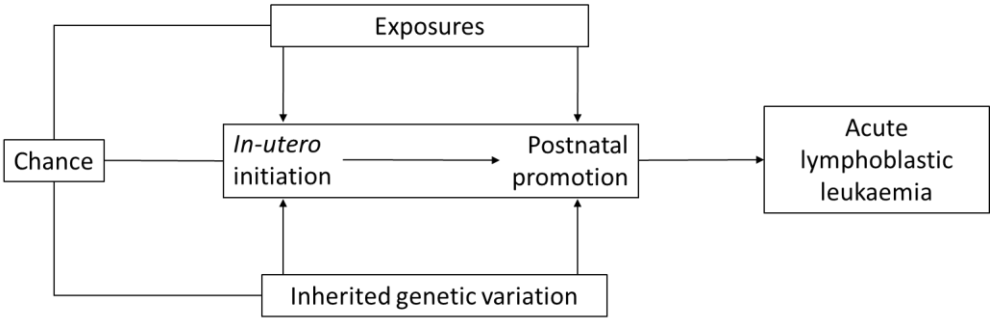


Figure 4. Causality of childhood acute lymphoblastic leukaemia. (Inaba, Greaves and Mullighan, 2013).

An alternative yet similar hypothesis of disease initiation is the "delayed infections hypothesis" (Greaves, 2018). Figure 5 illustrates this hypothesis in which a "first hit" is acquired *in utero* and limited exposure to bacterial or viral infections in the first year of life (likely due to lack of attendance of a day care or exposure to pets) results in an improperly developed immune system which proliferates subsequent mutations that lead to overt ALL.

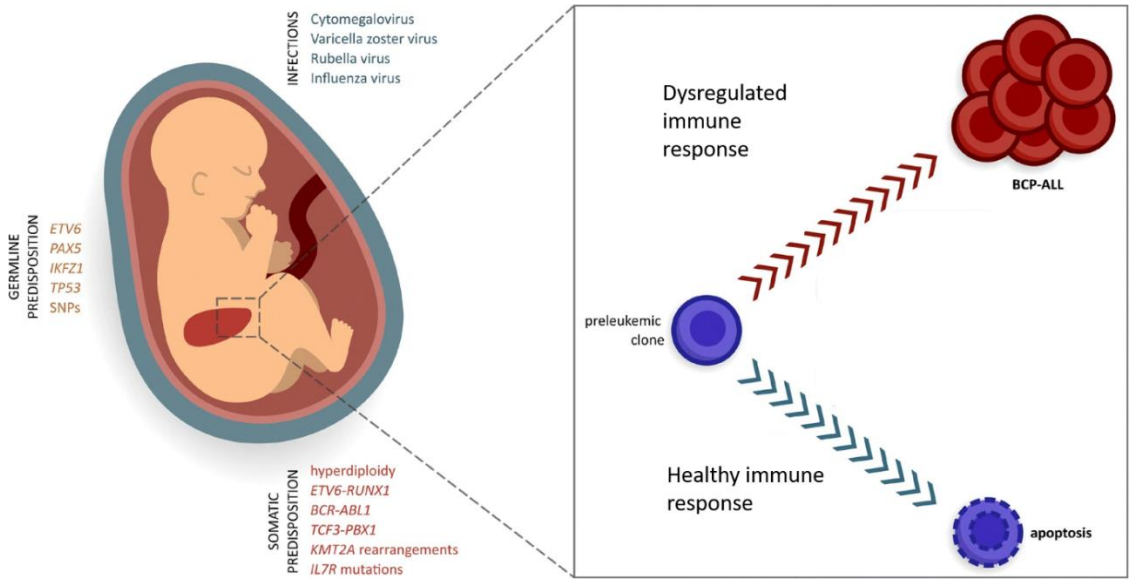


Figure 5. Illustration of the development of acute lymphoblastic leukaemia through infection. Germline mutations form the primary hit which requires a second hit of a somatic mutation developed due to the abnormal response of the immune system. (Rüchel *et al.*, 2022).

Risk factors theorised to influence the development of a “first hit” or impart the necessary “second hit” can be both intrinsic and extrinsic. Intrinsic factors thought to increase the risk of developing ALL include increased maternal age, high birth weight, pre-labour caesarean delivery, having a twin, having Down syndrome or Noonan syndrome, and maternal diabetes (Williams *et al.*, 2019; Søegaard *et al.*, 2018; Tebbi, 2021; Rüchel *et al.*, 2022; Onyije *et al.*, 2022). Whilst extrinsic factors increasing risk include parental exposure to pesticides or paint, increased maternal coffee intake, limited or absent microbial exposures in earlier life, and exposure to ionising radiation (Greaves, 2018; Williams *et al.*, 2019). Conversely, there are factors associated with a reduced risk of developing childhood ALL such as breastfeeding, maternal intake of folic acid, day-care attendance, contact with dogs or cats in the first year of life, and the presence of allergies (Williams *et al.*, 2019; Onyije *et al.*, 2022).

1.3.3 Epidemiology

ALL is the most common form of childhood cancer, accounting for approximately 75-80% of acute leukaemia in this age group whilst leukaemia accounts for roughly a third of all childhood cancer (Lustosa de Sousa *et al.*, 2015). In contrast to most cancer types, ALL is rare in adults with the lowest incidence rates being reached at age 30-34 in females and 45-49 in males and around 60% of diagnoses occurring before 20 years old (Malard and Mohty, 2020). Furthermore, only ~5 in 100 new cases occur in people over 75 years old (Cancer Research UK, 2024a). The incidence of ALL by age is bimodal, with a peak prevalence between the ages of 2 and 5 years old and a second, smaller peak occurring after the age of 65 as shown in Figure 6. There is a disparity in incidence of ALL by sex with a ratio of 1.4 males to every 1 female, with males also having a greater risk of relapse and death (Moorman, 2016).

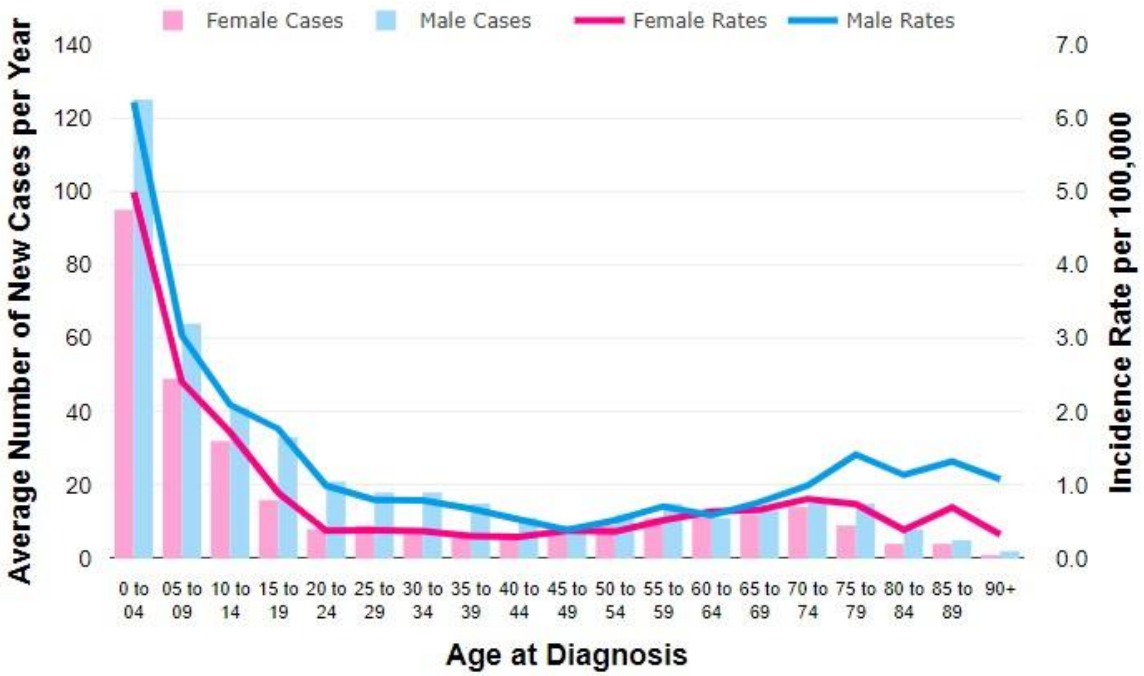


Figure 6. Incidence of acute lymphoblastic leukaemia by age in the UK. (Cancer Research UK, 2024a).

Before the 1960's, ALL was considered practically untreatable with an almost 100% rate of fatality (Barnes, 2008). In recent years, survival rates of paediatric ALL have improved significantly to approximately 90% (Malard and Mohty, 2020). This improvement over the decades is depicted in Figure 7. This progress is due to the adoption of multiagent chemotherapy regimens (Section 1.4), risk stratification of patients according to clinical and biological features as well as early treatment response to assign varying treatment intensity, (Section 1.6) and also better treatment/ control of infections (Hunger and Mullighan, 2015). However, despite high cure rates, relapsed ALL remains a leading cause of morbidity and mortality in children (Mullighan, 2012).

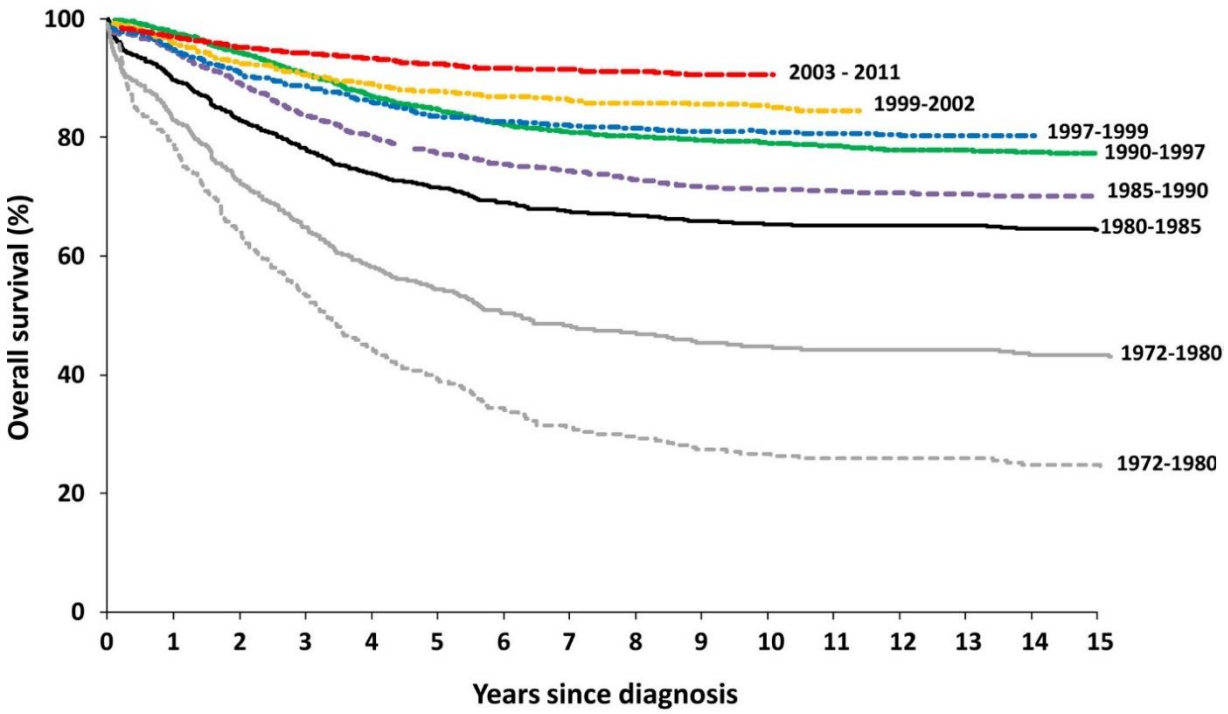


Figure 7. Kaplan Meier illustrating the improvement in survival for successive UK paediatric ALL trials. (Bartram, Veys and Vora, 2020).

In contrast to the impressive survival rates of paediatric ALL, outcome for adults is comparatively dismal with cure rates of less than 40% (Samra *et al.*, 2020). Outcome gets progressively worse with age with outcome for people over 50 falling to ~25% after 5 years (Malard and Mohty, 2020). This deterioration in survival with age is depicted in Figure 8. The disparity between survival of paediatric and adult patients has been attributed to a multitude of factors including differences in disease cytogenetics and genomic landscapes, an increased frequency of T-cell ALL, and children's ability to withstand much more intensive chemotherapy due in part to organ dysfunctions and comorbidities in older patients (Neaga *et al.*, 2021). Only paediatric ALL will be considered in this thesis.

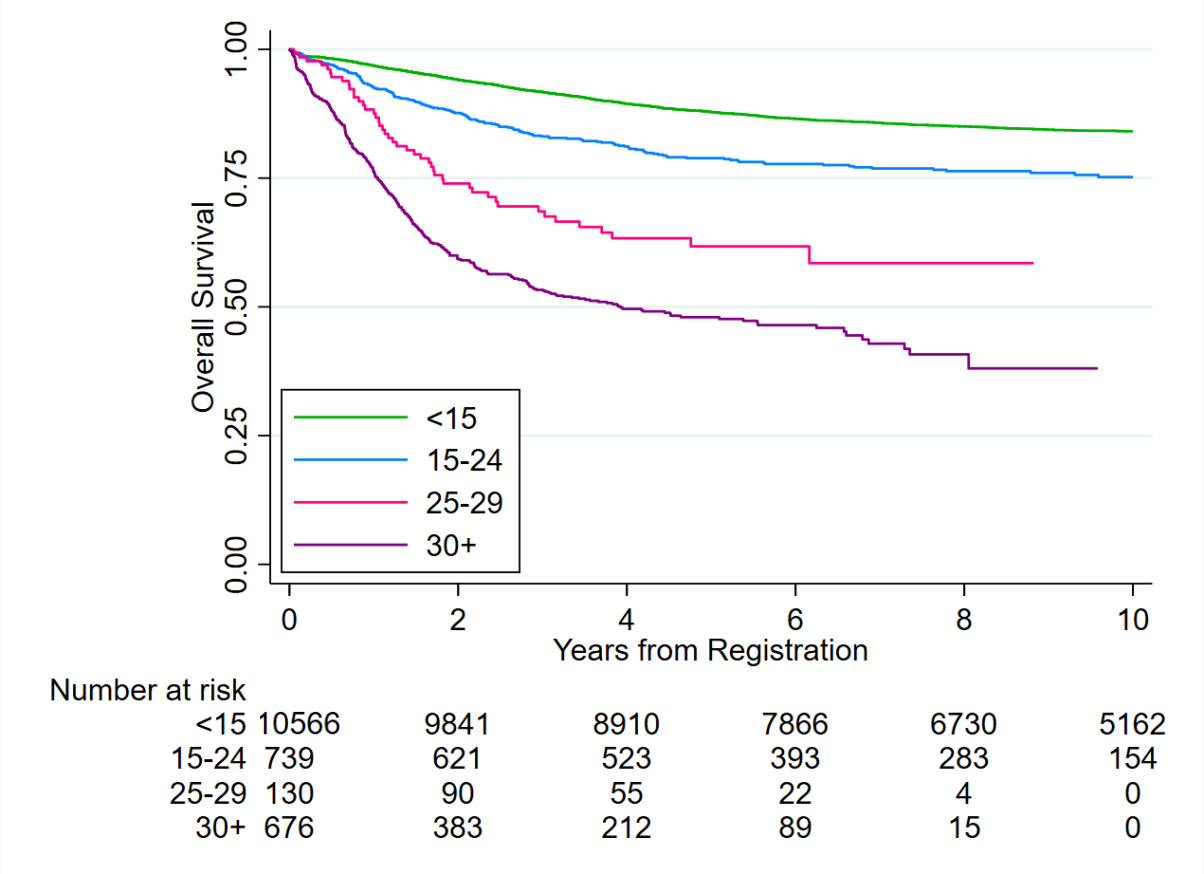


Figure 8. Kaplan Meier depicting difference in survival by age group from patients enrolled on paediatric trials UKALLXI – UKALL2011 and adult trial UKALL60+.

1.3.4 Subtypes

Determination of ALL subtypes is essential for the correct diagnosis and treatment of patients. Immunophenotyping remains the gold standard for identifying lineage of leukaemic lymphoblasts with T-cell, mature B-cell, and B-cell precursor phenotypes demonstrating immense therapeutic importance (Pui, Robison and Look, 2008; Malard and Mohty, 2020). Chromosomal analysis is an integral part of initial work-up of acute lymphoblastic leukaemia which was historically performed through conventional cytogenetics, fluorescence in-situ hybridisation (FISH), flow cytometry, and real-time polymerase chain reaction (RT-PCR) to determine cytogenetic features for prognosis and risk stratification (Malard and Mohty, 2020; Pui, Robison and Look, 2008). Presently, SNP-array, RNA fusion panel, and whole genome sequencing are more regularly used.

1.3.4.1 Immunophenotype

Patients can be broadly categorised as B- and T-lineage ALL with B-lineage ALL accounting for ~85% of childhood cases (Heim and Mitelman, 2009). Early studies of B-lineage ALL identified a difference in outcome by subtypes of B-ALL, however, improved treatment approaches as well as recognition of genetic abnormalities has lessened the importance of these classifications (Pui, 2012). 5-year event-free survival and overall survival rates for B-cell ALL patients are above 85% and 95% respectively on modern childhood clinical trials, whilst T-cell ALL outcomes are inferior by ~5-10% (Teachey and Pui, 2019). This disparity can be attributed to a multitude of factors including an older population in T-cell cohorts, poorer tolerance and increased resistance to chemotherapy agents, a lower proportion of favourable genetics, and an increased risk of extramedullary relapse (Teachey and Pui, 2019). To alleviate this difference, more intensive chemotherapy is generally prescribed to T-cell patients, while a recent emphasis on identifying genetic insights and alternative therapies has been placed (Van Vlierberghe *et al.*, 2008; Coustan-Smith *et al.*, 2009).

1.3.4.2 Genetics

The focus of this thesis will be on the good risk genetic subtypes *ETV6*::*RUNX1* and high hyperdiploidy with the rationale that due to outstanding survival rates in these groups, there exists optimal treatment elements for these patients, which maintain impressive outcomes whilst minimising toxicities and long-term late effects. Furthermore, these two subtypes account for >50% of all childhood ALL cases, thus efforts to reduce the late-effects in these groups would benefit the majority of childhood ALL survivors.

1.3.4.2.1 ETV6::RUNX1

The *ETV6*::*RUNX1* fusion is one of most common genetic aberrations in childhood ALL occurring in about 25% of B-precursor cases (Forestier *et al.*, 2008). This genetic abnormality is a translocation of a 12 and 21 chromosome resulting in t(12;21)(p13;q22) which can be detected by FISH or RT-PCR (Moorman, 2016). *ETV6*::*RUNX1* is associated with other good risk prognostic features such as low white cell count and age (Enshaei *et al.*, 2013). These patients are generally treated on standard or low risk protocols. *ETV6*::*RUNX1* patients have an excellent prognosis with survival rates surpassing 95% at 5-years on contemporary paediatric trials [Figure 7, Table 1]. This exceptional outcome is seen across multiple treatment protocols from national or collaborative group clinical trials (Østergaard *et al.*,

2024). Due to this, *ETV6*::*RUNX1* patients are now being considered for treatment de-escalation on future protocols (Austin and Patel, 2023; Gandemer *et al.*, 2009; Monovich, Gurumurthy and Ryan, 2024).

		5-year survival rates		
Study Group	**Clinical Trial**	**Overall**	**Event-free**	**Relapse rate**
CCLSG	ALL2004	96%	92%	6%
CoALL	07-03	98%	92%	6%
NOPHO	ALL2008	98%	92%	6%
UKALL	2003	96%	94%	4%
AIEOP-BFM	ALL 2000	98%	93%	6%
ALL-IC BFM	ALL 2002	94%	87%	12%
DCOG	ALL10	98%	95%	5%
JACLS	ALL-02	98%	95%	5%

Table 1. Five-year survival rates of *ETV6*::*RUNX1* patients across study groups. CCLSG Children's Cancer and Leukemia Study Group, CoALL Childhood Acute Lymphoblastic Leukemia, NOPHO Nordic Society of Pediatric Hematology and Oncology, UKALL United Kingdom Acute Lymphoblastic Leukaemia, AIEOP-BFM Associazione Italiana di Ematologia e Oncologia Pediatrica and Berlin Frankfurt Münster, ALL IC-BFM Acute Lymphoblastic Leukemia Intercontinental-Berlin Frankfurt Münster, DCOG Dutch Childhood Oncology Group, JACLS Japan Childhood Leukemia Study Group. Data source: (Østergaard *et al.*, 2024).

1.3.4.2.1 High Hyperdiploidy

High hyperdiploidy (HeH) is the most common cytogenetic abnormality occurring in ~25-30% of childhood B-cell precursor ALL patients (Woodward *et al.*, 2023). HeH is characterised as the non-random gain of chromosomes X, 4, 6, 10, 14, 17, 18, and 21 resulting in a minimum of 51 chromosomes (Paulsson and Johansson, 2009). There is variation in the definition of high hyperdiploidy amongst study groups, with a general consensus definition of aneuploidy between 51-65 or 67 chromosomes (Enshaei *et al.*, 2021; Teachey and Pui, 2019; Haas and Borkhardt, 2022; Moorman *et al.*, 2003). High hyperdiploidy is associated with a good outcome having >90% 5-year survival rates and is generally associated with other good risk factors such as low white cell counts and a peak incidence between the age of 2-4 years old

(Paulsson, 2015). Despite this excellent prognosis, ~10-15% of high hyperdiploidy patients relapse accounting for up to 25% of all childhood ALL relapses due to the group's prevalence and the wide biological heterogeneity across hyperdiploidy (Paulsson *et al.*, 2010; Irving *et al.*, 2016; Lee *et al.*, 2023a; Moorman *et al.*, 2022a). Thus, studies have endeavoured to identify risk factors within this subtype for further stratification resulting in the favourable risk groups of triple trisomy, double trisomy, a DNA index between 1.16-1.6, and UKALL low-risk high hyperdiploidy profile described in Table 2 (Reismüller *et al.*, 2017; Enshaei *et al.*, 2023; Lee *et al.*, 2023a; Chilton *et al.*, 2014; Sharathkumar *et al.*, 2008; Aricò *et al.*, 2008).

Subgroup	**Definition**
Triple trisomy (TT)	Simultaneous trisomy of chromosome 4, 10, and 17
Double trisomy	Simultaneous trisomy of chromosome 4 and 10
UKALL low-risk high hyperdiploidy profile	EITHER simultaneous trisomy of 17 and 18 OR +17 or +18 in the absence of +5 and +20.
DNA index	DNA index between 1.16-1.6

Table 2. Favourable high hyperdiploidy subgroups.

1.4 Treatment

Current treatment regimens for ALL are divided into five stages: induction therapy, CNS-directed treatment, consolidation, delayed intensification, and maintenance therapy as seen in Figure 9 below.

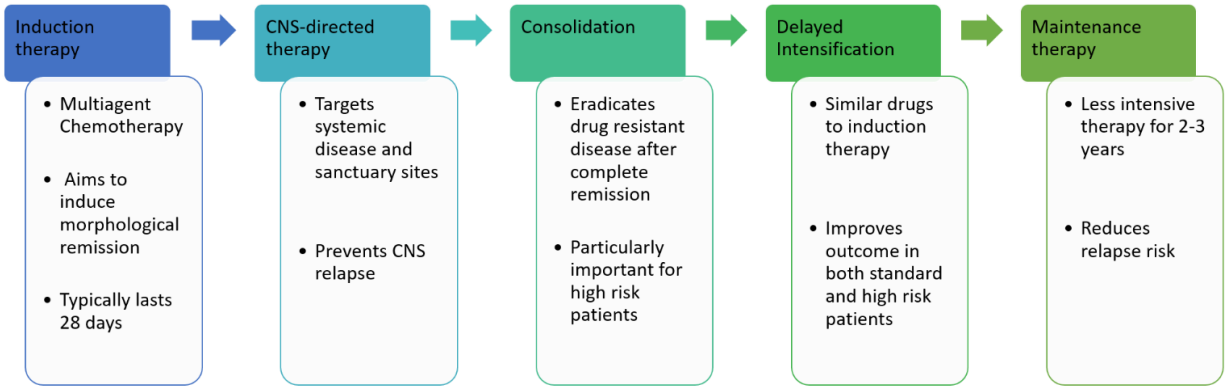


Figure 9. Exemplar treatment regimen for paediatric ALL.

1.4.1 Induction

Induction treatment is a multiagent chemotherapy regimen with the aim of inducing morphologic remission and restoring normal bone marrow haematopoiesis. The drugs given during induction often include vincristine, corticosteroids, namely dexamethasone or prednisolone, and asparaginase. Some regimens will also add an anthracycline, usually doxorubicin, or daunorubicin, to higher risk group treatment arms. The expectation is that patients will be in complete remission (CR) (defined as <5% blasts in the bone marrow by morphology) within 28 days of the start of treatment (Buchmann *et al.*, 2022). Most paediatric patients (95%) on modern protocols will achieve this benchmark (Cooper and Brown, 2014). Approximately half of patients who fail to attain CR within the first 4 weeks will endure toxicity resulting in treatment-related mortality (Schrappe *et al.*, 2012). The remaining half will have resistant disease. For these patients, an allogeneic bone marrow transplant is usually pursued.

1.4.2 CNS-directed therapy

Treatment of ALL targets both systemic disease and sanctuary sites – extramedullary anatomic locations that are difficult to penetrate with systemic chemotherapy alone. The most important of these sanctuary sites is the central nervous system (CNS) (Laningham *et al.*, 2007). Whilst only 3% of patients have detectable CNS involvement at diagnosis, 50%-70% of patients will develop CNS leukaemia unless specific CNS-directed therapy is given (Seibel,

2008). Thus targeting this sanctuary site is critical in order to prevent CNS relapse in paediatric patients.

Cranial irradiation is limited to use in patients who are at high risk of CNS relapse due to the associated extreme rates of neurotoxicity (Pui, 2006). Within an intergroup collaboration of 10 major childhood treatment groups, it was found that the proportion of newly diagnosed patients assigned to receive cranial irradiation ranged from 0%-33% (Vora *et al.*, 2016). In current UKALL trials, only patients with a CNS3 status (a non-traumatic cerebrospinal fluid sample that contains $\geq$ 5 WBC/μL with identifiable blasts, or presence of a cerebral mass or cranial palsy) receive cranial irradiation which amounts to <5% of patients. It was found that among high-risk patients without a CNS3 status, that a radiation dose of 12 Gy (in place of the full 18 Gy dose) can be given without increased risk of CNS relapse, further decreasing the risk of neurotoxicites (Schrappe *et al.*, 2000; Riehm, 1991). The full 18 Gy dose is administered to patients with CNS leukaemia at diagnosis (Pui, Robison and Look, 2008). Several paediatric trials have tested the effects of omitting cranial irradiation from treatment altogether and found that there was no increased risk of CNS or systemic relapse (Vilmer *et al.*, 2000).

For the remainder of patients, direct intrathecal (IT) administration of chemotherapy and systemic administration of chemotherapy able to penetrate the blood-brain barrier are used to eradicate CNS disease (Cooper and Brown, 2014). This therapy usually takes the form of weekly or bi-weekly IT therapy with systemic drugs such as dexamethasone, asparaginase, 6-mercaptopurine, cytarabine and high-dose methotrexate (HD-MTX) (Seibel, 2008). Options for intrathecal chemotherapy include including intrathecal methotrexate (IT-MTX) or a combination of IT-MTX, cytarabine, and hydrocortisone (Cooper and Brown, 2014).

1.4.3 Consolidation and intensification

The purpose of consolidation therapy is to eradicate any remaining drug-resistant residual disease after complete remission is achieved (Pui, Robison and Look, 2008). The agents used in this stage of therapy are combinations to maximise synergy and minimise drug resistance and are often agents not used during induction, such as high-dose asparaginase (25,000 IU/m^2) given for an extended period, HD-MTX with mercaptopurine, thioguanine, and cytarabine (Pui and Evans William, 2006; Cooper and Brown, 2014; Pession *et al.*, 2005; Amylon *et al.*, 1999; Sallan *et al.*, 1983; Silverman *et al.*, 2001). Reinduction treatment may

also be used on certain regimens in which patients are treated with the same agent given during induction (Pui, Robison and Look, 2008). For patients with high-risk ALL, both of these treatments may be used as these children often receive more intensive consolidation regimens over a longer period (Cooper and Brown, 2014). Consolidation therapy is particularly important in high-risk patients such as T-cell ALL, infant ALL, as well as patients with overt CNS disease (Esparza and Sakamoto, 2005).

1.4.4 Maintenance therapy

Previous studies have shown that children with ALL need continued treatment, at a much less intensive rate, in order to drastically reduce the risk of relapse (Pui, Robison and Look, 2008; Hunger and Mullighan, 2015; Cooper and Brown, 2014; Pui and Evans William, 2006). It is believed that this stage of therapy is necessary in order to eradicate residual leukaemic cells that are cycling very slowly and are essentially chemo-resistant when not dividing. Thus, the low levels of chemotherapy over a long period kill these cells when they eventually move into the dividing phase. Historically, maintenance therapy often lasted 2 years for girls with an additional year (3 years total) of therapy for boys due to higher incidence of relapse and lower survival rates (Gupta *et al.*, 2022). However, recent studies have demonstrated that this additional year of therapy was not beneficial, and now boys and girls receive the same length of maintenance therapy on most modern protocols. The backbone of current regimens consists of daily mercaptopurine and weekly methotrexate dosages (Pui, Robison and Look, 2008; Cooper and Brown, 2014; Pui and Evans William, 2006). In some protocols, additional pulses of vincristine and corticosteroids are added (Seibel, 2008).

1.4.5 Adverse effects from treatment

Whilst current protocols comprising multi-agent chemotherapy have led to impressive survival rates in paediatric ALL, they are also the cause of many severe adverse effects including pancreatitis, thromboembolism, neuropathy, and endocrinopathies such as pituitary deficits and hypothyroidism (Lejman *et al.*, 2021; Zhang and Gu, 2023; Follin, 2019; Howard and Pui, 2002). Approximately 50% of all patients will be affected by at least one of 14 common severe acute toxic effects of chemotherapy (Schmiegelow *et al.*, 2016). Furthermore, many cancer survivors suffer from long term sequelae including chronic health disorders and neurobehavioral problems (Kato and Manabe, 2018). The prevalence of these adverse effects varies in long term survivors of paediatric ALL based on the type of therapy

given as well as relapse status with non-irradiated and non-relapse patients reportedly having fewer chronic health conditions. Patients who received therapeutic irradiation are at the highest risk for secondary neoplasms which, along with recurrence of ALL, is a major cause of excess mortality in survivors (Mody *et al.*, 2008).

Many of the drugs given during therapy are known to cause adverse effects (Langebrake, Reinhardt and Ritter, 2002). Examples of this include anthracycline which has been found to cause chronic cardiotoxicity, high-dose methotrexate which can cause nephrotoxicity, and asparaginase which, among other things, has been associated with development of pancreatitis and hypersensitivity to the drug (Schmiegelow *et al.*, 2016; Langebrake, Reinhardt and Ritter, 2002; Childhood Acute Lymphoblastic Leukaemia Collaborative Group, 2009). A more detailed view of the adverse effects of chemotherapy agents is detailed in Table 3. These effects have been found to correlate with the administered dose (Langebrake, Reinhardt and Ritter, 2002). As such, increases in treatment intensification over time have resulted in treatment-related death being just as likely as leukaemic relapse in low-risk patients (Schmiegelow *et al.*, 2016). In consequence, the focus of much research has been to identify optimal dosages of therapeutic drugs to ensure cure with minimal toxicity as well as to eradicate the need for irradiation except for very high risk cases (Pui and Howard, 2008; Essig *et al.*, 2014) (Möricke *et al.*, 2008). Furthermore, there has been a surge in the development of alternate drugs and targeted therapies to lessen the therapeutic burden of conventional chemotherapy whilst still maintaining excellent outcomes (Lejman *et al.*, 2021).

Chemotherapy Agent/Therapy	Adverse Effect(s)
Anthracyclines	Cardiotoxicity, Myelosuppression, Hair loss, Mucositis, Nausea, Vomiting, Chemical cellulites
L-Asparaginase	Allergy, Pancreatitis, Thrombosis, Hyperglycemia, Encephalopathy, Hepatodysfunction, Nausea, Vomiting
Cranial Irradiation	Post-treatment somnolence, Seizures, Brain tumours, Hair loss, Osteoporosis, Neurological/endocrine dysfunction, Secondary Neoplasm, Learning deficits, Growth Hormone deficiency, Obesity, Stroke, Dental problems
Cytarabine	Arachnoiditis, Neurotoxicity, Headaches, Nausea, Vomiting, Fever, Rashes, Decreased fertility, Mucositis, Hepatodysfunction, Conjunctivitis
Cyclophosphamide	Myelosuppression, Bladder cancer, Acute myelogenous leukaemia, Decreased fertility, Haemorrhagic cystitis, Hair loss, Nausea, Vomiting, Increased antidiuretic hormone secretion
Etoposide	Myelosuppression, Allergy, Acute myelogenous leukaemia, Mucositis, Hair loss, Nausea, Vomiting
Glucocorticoids	Sepsis, Infection, Bone toxicities, Osteopenia, Osteonecrosis, Steroid psychosis, Mood disorders, Proximal myopathy, Acne, Hyperglycaemia, Hypertension, Weight gain
Mercaptopurine	Osteoporosis, Myelosuppression, Sun sensitivity, Hepatodysfunction, Mucositis, Nausea, Vomiting
Methotrexate	Nausea, Vomiting, Liver dysfunction, Myelosuppression, Osteoporosis, Sun sensitivity, Leukoencephalopathy Intrathecal administration adverse effects: Headaches, Fever, Seizures
Vincristine	Peripheral neuropathy, Pain, Constipation, Seizures, Hair loss, Chemical cellulites

Table 3. Adverse effects of chemotherapy agents and therapies. Sources: (Redaelli *et al.*, 2005; Inaba and Pui, 2010; Pui *et al.*, 2003; Oeffinger and Hudson, 2004; Pui and Jeha, 2007; Childhood Acute Lymphoblastic Leukaemia Collaborative Group, 2009).

1.5 Outcome

Outcome of childhood acute lymphoblastic leukaemia is outstanding at >90% overall survival and ~85% event-free survival at 5 years. This is primarily due to the prevention of relapse in the form of improvements in treatment and risk stratification as well as the addition of targeted therapies for certain subtypes (Bailey *et al.*, 2008). A major historic cause of poor outcome was the prevalence of CNS relapse largely resulting in death. However, studies in the 1960s and 1970s identified that administering CNS-directed therapy to all patients was necessary for cure (Thastrup *et al.*, 2022). This resulted in the incidence of CNS relapse falling from >50% of patients to less than 10%, with rates on modern protocols reported as low as 3% (Eden, 1995; Pui and Howard, 2008). The introduction of this phase of therapy initially consisted of cranial or craniospinal irradiation, but due to evidence of long-term sequelae of this form of treatment, it has been largely replaced by intrathecal and systemic chemotherapy (Richards *et al.*, 2013).

Following the success in the reduction of CNS relapse, focus has shifted to marrow relapses which despite improvement over the last few decades, has a recurrence rate of 10-15% (Bailey *et al.*, 2008). This recurrence is often salvageable however with 10% of early marrow relapses and 50% of late marrow relapses being cured (Bailey *et al.*, 2008). Prognosis for children who relapse is poor primarily in high risk groups who have a chance of survival as low as 25% (Irving, 2016). This highlights the importance of treatment reduction only in good risk patients to maintain low relapse rates in paediatric ALL patients. The development of new targeted therapies in recent years has improved outcomes in patients with relapse B-cell ALL, however, suggesting a promising future for the treatment of relapsed ALL (Malard and Mohty, 2020). Nowadays, remission death is as common as relapse highlighting the major advancements in relapse prevention over the previous decades.

1.6 Prognostic factors and risk stratification

There are a host of prognostic factors in paediatric ALL that are used for risk stratification. Patient information at the point of diagnosis including a patients' age and initial white cell count are generally used to determine initial treatment arms with older children and children with a high WBC count at diagnosis being classified as high risk (Hayashi, Makimoto and Yuza, 2024). Response to treatment, usually measured by a patient's measurable residual disease (MRD) level at the end of induction (EOI) and ability to achieve a complete remission, is then

used to adjust risk groups as necessary (Borowitz *et al.*, 2008; Conter *et al.*, 2010). Genetics are also of great consideration when it comes to ascertaining patient risk (Hayashi, Makimoto and Yuza, 2024).

1.6.1 Age

Age is an independent predictor of outcome (Seibel, 2008). Patients aged between 1 and 9 years (approximately 75% of childhood ALL patients) have the best prognosis with a higher disease-free survival rate than any other age group. Favourable prognostic factors are also more common in this age group, with good risk cytogenetics and B-cell disease a more frequent occurrence. Generally, survival is directly correlated with patient age with risk increasing as age increases. The exception to this is in infants <1 years who have relatively poor survival in comparison to children, with an EFS of 47% at four years (Lee and Cho, 2017). Some hypothesise that the prognostic effect of age is a surrogate for other factors such as older patients' inability to tolerate more intensive treatment and a greater prevalence of high-risk abnormalities such as *BCR*::*ABL1*. Similarly, it is suggested the poor outcome in infants can be attributed to *KMT2A* rearrangements that are associated with hyperleukocytosis as ~80% of infants have this subtype (Malard and Mohty, 2020).

1.6.2 Sex

In paediatric ALL, male patients have been shown to have a worse outcome than females (Gupta *et al.*, 2022). Many have hypothesised this is due to the impact of testicular relapse, however this is unclear as some studies postulate that increased rates of central nervous system (CNS) relapse are the cause, whilst others find that gender bears no significant difference on outcome (Gupta *et al.*, 2022; Friedmann and Weinstein, 2000). Male patients are still given additional maintenance therapy on some protocols, however many modern protocols no longer have this distinction due to evidence that longer therapy for boys has no impact on outcome, which is instead dependent on the backbone and intensity of chemotherapy administered before maintenance (Teachey, Hunger and Loh, 2021).

1.6.3 Central nervous system

The CNS is a frequently affected site at diagnosis (<5%) and at relapse (up to 30-40%) (Cancela *et al.*, 2012). Before the introduction of prophylactic CNS irradiation, ~75% of relapse involved the CNS (Kreuger *et al.*, 1991). Thus, irradiation was considered vital in the cure of ALL and

was once considered the standard. However, long term effects of irradiation have been reported including late neurocognitive deficits, secondary cancers, and excess late mortality. As such, current treatment involves serial intrathecal chemotherapy in conjunction with methotrexate as an alternative to irradiation in most or all paediatric patients (Malard and Mohty, 2020). CNS irradiation is generally now reserved for patients with CNS involvement at diagnosis or those considered high risk of a CNS relapse, although several studies have shown that only patients who present with CNS disease at the time of diagnosis (~2%) benefit from cranial radiotherapy (CRT) and that the use of CRT in first-line therapy for other patients did not affect outcome (Vora *et al.*, 2016). It has been found that children who have a late ($\geq$ 18 months from end of treatment), isolated CNS relapse and did not receive cranial irradiation often have a great outcome with a 5-year EFS similar to newly diagnosed patients. However, patients who relapse in the CNS shortly after remission or who received cranial irradiation have very poor outcomes and, as such, this is a very poor prognostic indicator (Cancela *et al.*, 2012; Pui, 2006).

1.6.4 White blood cell count

The initial WBC count at diagnosis is a well-studied and oft reported prognostic factor in ALL (Hastings *et al.*, 2014). There is a strong association between outcome and white cell count with increased risk of death as initial white cell count at diagnosis increases. (Hastings *et al.*, 2014). Patients with a WBC count >50 x 10^9 for B-cell were found to have an increased risk regarding both overall survival (OS) and disease-free survival (DFS) and so this is often used as a cut-off in risk stratification (Eden *et al.*, 2000). Whilst many prognostic factors have lost significance as treatment has improved, white cell count continues to be a strong prognostic indicator of outcome both individually and in combination with other features (Pui and Evans William, 2006).

1.6.5 Immunophenotype

T-cell ALL accounts for around 15% of childhood ALL cases (Shuster *et al.*, 1990). Historically, T-cell ALL patients have lower survival rates than their B-Cell precursor counterparts (Goldberg *et al.*, 2003). However, on current treatment protocols, T-cell patients are generally treated with a more intensive regimen resulting in outcomes like that of children with B-ALL, with survival rates of approximately 80% (Coustan-Smith *et al.*, 2009). Paediatric B-ALL

patients have excellent outcomes with ~90% achieving long term remission (Moorman *et al.*, 2022a).

1.6.6 Cytogenetics

Acquired chromosomal abnormalities are closely associated with the biology of ALL and indicate the genes involved in leukaemogenesis (Harrison, 2001). As such, paediatric patients are often stratified by genetic subtype due to their prognostic ability. Genetic abnormalities can be considered as primary or secondary events where primary abnormalities cause the initiation of a pre-leukaemic clone and secondary abnormalities converts the clone into overt ALL (Moorman, 2016). Generally, primary abnormalities are chromosomal translocations whilst secondary abnormalities comprise of copy number alterations and point mutations. Whilst there are abnormalities that have been reported as both primary and secondary in different settings, generally they are distinct groups. Most risk stratification methods focus on primary abnormalities as these have a stable prognostic effect. An overview of these primary abnormalities is given in Table 4.

Subtype
t(9;22)(q34.1;q11.2)/ *BCR*::*ABL1*
KMT2A rearranged
t(12;21)(p13.2;q22.1)/ *ETV6*::*RUNX1*
High Hyperdiploidy
Low Hypodiploid
Near haploid
t(5;14)(q31.1;q32.3)/ *IL3*::*IGH*
t(1;19)(q23.3;p13.3)/ *TCF3*::*PBX1*
BCR::*ABL1*-like, ABL-1 class fusion
BCR::*ABL1*-like, JAK-STAT activated
BCR::*ABL1*-like, NOS
iAMP21
MYC rearrangement
DUX4 rearrangement
MEF2D rearrangement
ZNF384 rearrangement
NUTM1 rearrangement
HLF rearrangement
UBTF::*ATXN7L3/PAN3,CDX2 ("CDX2/UBTF")*
IKZF1 N159Y
PAX5 P80R

Table 4. Subtypes of B-cell precursor ALL defined in the International Consensus Classification. Information source: (Duffield, Mullighan and Borowitz, 2023).

In several classifications, including the ones used for the data in this thesis, *ETV6*::*RUNX1* and high hyperdiploidy make up the good risk cytogenetic risk group whilst others such as a *BCR*::*ABL1* fusion (t(9;22)), intrachromosomal amplification of chromosome 21q (iAMP21), and rearrangements of the *KMT2A* gene are associated with a poor prognosis and, along with near haploidy, low hypodiploidy, and t(17;19), make up the high risk cytogenetic group. All remaining abnormalities fall into the intermediate risk category (Hakeem *et al.*, 2014;

Chennamaneni *et al.*, 2018; Moorman *et al.*, 2010). Whilst many children with B-ALL harbour one of the major chromosomal abnormalities, the remaining patients (~30%) are classified as B-other ALL and are categorised as intermediate risk (Schwab *et al.*, 2022). However, recent studies have shown that approximately one third of B-other-ALL patients can be classified into a distinct genetic subgroup with their own prognosis. Two examples of this are ABL-class fusions (~4% of B-other-ALL) which are associated with a high risk of relapse and *ERG* deletions (10%-15% of B-other-ALL) that had 10-year EFS rates of 97.2%, indicating patients with this abnormality have an excellent prognosis (Schwab *et al.*, 2022; Moorman, 2016). An overview of the main genetic abnormalities is depicted in Table 5. The prognosis of these genetic abnormalities is well reported in the literature and is summarised by Roberts *at al.* (Roberts, 2018).

Genetic Abnormality	Description	Prognosis	Prevalence
High Hyperdiploidy	51-67 chromosomes	Good (>90%)	25-30%
t(12;21)(p13;q22)	*ETV6::RUNX1* / *TEL::AML1* fusion	Good (>90%)	25%
t(1;19)(q23;p13)	*TCF3::PBX1* fusion	Good with modern intensive protocols (>80%)	~5%
KMT2A	11q23 rearrangements, *KMT2A* rearranged, t(4;11)(q21;q23), t(11;19)(q23;p13.3), t(9;11) del(11)q23	Poor (<75%)	~5%
t(9;22)(q34;q11.2)	*BCR::ABL1* fusion gene, Philadelphia positive	Poor (<75%)	5%
iAMP21	Too many copies of a portion of chromosome 21	Poor (<75%)	~3%
Near Haploidy	24-29 chromosomes	Poor (<75%)	~2%
Low Hypodiploidy	30-39 chromosomes	Poor (<75%)	~1%
t(17;19)(q22;p13)	*TCF3::HLF* fusion	Poor (<75%)	1%

Table 5. Summary of the main genetic abnormalities in paediatric ALL.

The risk classification of the genetic abnormalities outlined above is the convention that will be used in this thesis. However, there is no standard definition for risk groups across different countries/research groups and as such, patients with the same genetic abnormality may be treated with different intensity therapies based on protocol definitions. A comparison of the different classifications of genetic abnormalities is shown in Table 6.

	Study Group						
Genetic Abnormality	**COG**	**St Jude**	**NOPHO**	**UKALL**	**AIEOP-BFM**	**DCOG**	**JACLS**
HeH	Favourable (DT)	Low	Any	Good	Standard/ Medium	Non-high/ High	Standard/ High
ETV6::*RUNX1*	Favourable	Low	Any	Good	Standard/ Medium	Non-high/ High	Standard/ High
TCF3::*PBX1*	NC	Standard	Inter/ High	Inter	Standard/ Medium	Non-high/ High	Standard/ High
KMT2A	Unfavourable	High (in infants)	High	Poor	High	High	High
BCR::*ABL1*	NE	High	NE	NE	High	High	High
iAMP21	Unfavourable	NC	Inter/ High	Poor	Standard/ Medium	Non-high/ High	Standard/ High
NH	Unfavourable	Standard	High	Poor	Standard/ Medium	Non-high/ High	Standard/ High
LH	Unfavourable	Standard	High	Poor	Standard/ Medium	Non-high/ High	Standard/ High
TCF3::*HLF*	NC	NC	Any	Poor	Standard/ Medium	Non-high/ High	Standard/ High

Table 6. Comparison of several genetic risk group classifications. Patients classified as more than one group or "any" were classified to groups based on other factors such as response to therapy or other clinical and demographic factors rather than just cytogenetics alone. COG: Children's Oncology Group, NOPHO: Nordic Society of Pediatric Hematology and Oncology, UKALL: United Kingdom Acute Lymphoblastic Leukaemia, AIEOP-BFM: Associazione Italiana di Ematologia e Oncologia Pediatrica and Berlin Frankfurt Münster, DCOG: Dutch Childhood Oncology Group, JACLS: Japan Childhood Leukemia Study Group. Inter: Intermediate, NC: Not Classified, DT: Double Trisomy, NE: Not Eligible.

1.6.7 Measurable residual disease

Measurable residual disease is the name given to the number of cells remaining in the body at the time of testing, for example at the end of induction (Leukaemia and Lymphoma Society, 2024). MRD testing is used to quantify these cells when the cancer is no longer detectable by blood tests (Chen *et al.*, 2024). Assessment of MRD is performed via three different techniques, namely polymerase chain reaction (PCR), flow cytometry, or next-generation sequencing (NGS) (Dekker *et al.*, 2023). The different techniques and their sensitivities are shown in Table 7.

Technique	**Sensitivity**
Flow cytometry	10^{-4} to 10^{-5} (0.01% to 0.001%)
Polymerase chain reaction	10^{-4} to 10^{-5} (0.01% to 0.001%)
Next-generation sequencing	10^{-6} (0.0001%)

Table 7. Techniques for assessing measurable residual disease. Information source: (Dekker *et al.*, 2023).

Measurable residual disease at the end of induction is a strong prognostic indictor and is often used for patient stratification in a clinical setting (Sutton *et al.*, 2009). Jacquy *et al.* (1997) demonstrated that a threshold of 0.001 (0.1%) for EOI MRD successfully identified a good risk group with 100% DFS and a poor risk group with 27% survival showing the prognostic value of MRD (Jacquy *et al.*, 1997). More recent studies, however, found a threshold of 0.01% to be most optimal with a hazard ratio of 0.18 (0.11-0.29) for MRD negative patients compared to MRD positive patients (Berry *et al.*, 2017). Children with MRD levels <0.01% are frequently assigned to non-high-risk groups due to their superior survival rates. However, the clinical significance of this threshold is unclear and there has been debate if this is the optimal threshold for prognosis (Sutton *et al.*, 2009; Jacquy *et al.*, 1997). In recent years, due to a lack of consensus on cutpoints and the loss of prognostic value associated with unnecessary categorisation, there has been a shift toward the use of continuous variables where possible (Donadieu *et al.*, 2000; Enshaei *et al.*, 2020).

Advances in patient risk stratification using early treatment response (measurement of MRD) and somatic genetic abnormalities have greatly improved treatment allocation and

nowadays, patients with low-risk disease are considered for treatment de-escalation while high-risk patients are allocated therapies that are more experimental. MRD and the majority of genetic abnormalities are prognostic rather than predictive biomarkers (Campana, 2010).

1.7 Risk-adapted therapy

Modern treatment of ALL in children and adolescents has seen an intensification for all patients contributing to the improvement in event-free survival (EFS) seen over the last 6 decades (Esparza and Sakamoto, 2005; Seibel, 2008). However, this approach has led to the overtreatment of some patients resulting in unnecessary toxicities (Seibel, 2008). This was the premise for "risk-adapted therapy" in which patients with better risk features are treated with less-intensive therapy whilst children with a lower probability of survival according to risk features receive more aggressive therapy (Cooper and Brown, 2014).

In 1993 at an international conference supported by the National Cancer Institute (NCI), a set of risk criteria were adopted by both the Pediatric Oncology Group (POG) and the Children's Cancer Group (CCG). The NCI criteria were based on known risk factors – age at diagnosis and initial white blood cell (WBC) count (Seibel, 2008). Risk classification has continued to evolve with the addition of other risk factors and the development of classifiers by other cooperative groups (Stanulla *et al.*, 2018; Moorman *et al.*, 2022b; Enshaei *et al.*, 2020).

In international paediatric ALL clinical trials, patients with good risk cytogenetics (*ETV6*::*RUNX1* and high hyperdiploidy), or patients with a good risk copy number alterations (CNA) profile coupled with good response to induction therapy will be assigned to treatment pathways which include dose reductions. In contrast, patients with high risk genetic abnormalities (*KMT2A* fusions, near haploidy, low hypodiploidy, intrachromosomal amplification of chromosome 21 (iAMP21) and *TCF3*::*HLF*) will be assigned to more intensive treatment irrespective of initial treatment response.

1.8 Differential outcome by genetics

It is clear within the literature that there is a differential outcome by genetics. This difference may be caused by underlying genetic-treatment interactions. There are two types of therapy to treat paediatric ALL – targeted and untargeted therapy. A clear example of a genetic-treatment interaction is that of the tyrosine kinase inhibitor (TKI) imatinib and *BCR*::*ABL1*. Imatinib is a targeted therapy that was developed to inhibit proliferation of cells transformed

by *BCR*::*ABL1* and induce apoptosis (Braun, Eide and Druker, 2020). Recent studies have shown that TKIs can also be used to treat ABL-class gene fusions (chimeric gene fusions whose functional consequence result in the constitutive activation of the ABL pathway, mimicking *BCR*::*ABL1* fusions) (Moorman *et al.*, 2020). Interactions between genetics and non-targeted therapy are much harder to explore and rely on *in vitro* or clinical evidence.

There is *in vitro* evidence for specific drug-genetic interactions in terms of sensitivity. For example, high hyperdiploid cells are particularly sensitive to methotrexate. Methotrexate polyglutamates have been shown to accumulate preferentially in lymphoblasts harbouring hyperdiploid compared with other genetic subtypes of ALL (Whitehead *et al.*, 1998). Cells harbouring an *ETV6*::*RUNX1* fusion have been shown to be particularly sensitive to asparagine treatment (Roel *et al.*, 2019). Asparaginase sensitivity was also shown to correlate with B cell differentiation (Huang *et al.*, 2024). It was also found that both *ETV6*::*RUNX1* and high hyperdiploidy are highly sensitive to prednisolone, vincristine, daunorubicin and L-asparaginase, four drugs commonly used in induction therapy, whilst poorer prognosis subtypes did not show that same sensitivity (Lee *et al.*, 2023b). Furthermore, patients with IK6 expression (aka *IKZF1* exon 4-7 deletion) were found to be sensitive to cytarabine and, conversely, showed resistance to both daunorubicin and asparaginase (Rogers *et al.*, 2021). Both methotrexate and asparaginase are mainstay drugs in ALL therapy. These in vitro observations are reflected by the treatment response of patients with these abnormalities. Patients with *ETV6*::*RUNX1* have a very rapid response to induction therapy that includes asparaginase and go on to have excellent long-term outcomes. In contrast, patients with high hyperdiploid have slower response to induction therapy (which does not include systemic methotrexate) yet still have excellent long-term outcomes. However, the clinical utility of this information is limited because no patient with ALL is treated with monotherapy. All patients with ALL are treated on multimodal chemotherapy regimens. Even those patients with *BCR*::*ABL1* or ABL-class fusion who receive targeted therapy with TKIs receive it as adjuvant chemotherapy rather than monotherapy.

There is also clinical evidence for drug-genetic interactions. Examples of these interactions in the literature include the Clappier *et al.* study which shows the adverse effect an *IKZF1* deletion harbours on event-free survival as well as the fact that vincristine-steroid pulses given to these patients during maintenance helps prevent relapses (Clappier *et al.*, 2015).

Furthermore, it was found that shorter maintenance therapy does not increase relapse risk for females or patients with either an *ETV6*::*RUNX1* or *TCF3*::*PBX1* fusion, but that the good prognosis reported for high hyperdiploidy patients highly depends on sufficient duration of maintenance therapy (Kato *et al.*, 2017). Currently there is little information about the interaction between somatic genetics and the response to specific therapeutic pathways, although some findings suggest that certain single-nucleotide-polymorphisms (SNPs) are associated with variations in therapy response (Yang *et al.*, 2009; Aplenc *et al.*, 2003; Ceppi *et al.*, 2014; Larkin *et al.*, 2023).

Whilst the interaction between genetics and response to therapeutic pathways could be relatively simple to identify and exemplify using traditional statistical methods such as a Cox regression model or regression techniques, it is likely that some of these interactions will be too complex to be fully explained in such simplistic terms. The future for risk stratification may lie in the use statistical models that account for the weighting of each factor as well as any significant interactions while assigning risk, rather than using prognostic factors individually (Enshaei *et al.*, 2020; Shouval *et al.*, 2021). This is now widely possible through the accessibility of large historic datasets and the use machine learning, a subdomain of artificial intelligence (AI), to assess data for models not interpretable using traditional statistical methods (Pan *et al.*, 2017).

1.9 Survival Analysis

Survival analysis is a collection of statistical methods which address questions regarding time to a specific event of interest (Guo, 2010). It is an area of research with many applications across areas such as engineering, biological, and social sciences (Klein and Goel, 2013). It can be used to investigate a variety of matters including recovery from a disability, the length of time children stay in foster homes, time between mental health treatment interventions, and the working life of a piece of machinery before repair is needed (Liu, 2012; Guo, 2010). Time to an event is the main interest of many cancer studies and as such, survival analysis is the most often used statistical method for analysing cancer clinical trial data (Clark *et al.*, 2003; Cox, 2022).

The time interval considered in survival analysis for cancer studies is typically from the day a patient starts treatment on a clinical trial to the time of last contact with the patient. If the

relevant event is not observed within this time interval then the survival time is considered right-censored (Jenkins, 2005). If an event is observed, then this length of time is known as the survival time of that individual (Pagano, Gauvreau and Mattie, 2022). In general, the most common events considered in survival analysis in a cancer setting are death, relapse, or secondary tumour with a patient considered to be in continuing remission (i.e. no event) otherwise. Traditional cross-sectional and longitudinal approaches to analysis are not widely applicable to time to event data because of censoring and the fact that the data are rarely Normally distributed with many events typically happening early and fewer later events occurring. (Clark *et al.*, 2003; Guo, 2010).

There are two functions which are the core of survival analysis, namely the survival function S(t) and the hazard function h(t) which are defined as the probability of surviving beyond a specified time point t and the probability of an event at time point t (or instantaneous rate of failure) respectively (Cox, 2022). The survivor function can be estimated nonparametrically using Kaplan-Meier methods (Kaplan and Meier, 1958) which can be used to quote survival rates at distinct time points (the percentage of patients who haven't had the event of interest at a given time point) and plot the estimated survival probability against time known as a Kaplan-Meier curve (Clark *et al.*, 2003). A comparison of the survival curves for 2 or more groups of interest is performed using the log-rank test to determine if the survival distribution of these groups is identical (Pagano, Gauvreau and Mattie, 2022). This is often utilised to determine information such as whether a specific treatment has had an impact on survival in one subgroup compared to those untreated or whether two different treatments have differing effect on survival for example.

It can also be of interest to study the relationship between the time to event outcome and variables such as demographic, clinical, and treatment factors which is done through Cox proportional hazards regression model (Cox, 1972; Dawson, Blanchette and Pihlstrom, 2021). This model calculates a hazard ratio (HR) for these factors from the hazard function and can be interpreted as the risk of an outcome of interest in one group (eg. males) / risk of an outcome of interest in another group (eg. females). For example, in a situation where the risk of death in men is twice that of the risk of death in women then the hazard ratio for men would be 2 compared to the baseline hazard group (denoted as 1) (Brody, 2012).

The use of survival analysis to assess treatment effect on patient outcomes is commonplace in childhood ALL research. Examples in the literature include Brown *et al.* and Locatelli *et al.* who utilise Kaplan Meier methods and Cox proportional hazards models to determine the effect of immunotherapy with blinatumomab vs chemotherapy on disease-free survival and event-free survival in relapsed patients respectively (Brown *et al.*, 2021; Locatelli *et al.*, 2021). For frontline therapy, Moorman *et al.* highlighted the benefit to overall, event-free and relapse-free survival of more intensive therapy for iAMP21 patients, whilst Vora *et al.* suggested that patients with higher levels of MRD at the end of induction had statistically higher event-free survival rate when treated with augmented post-remission therapy compared to standard therapy, but suffered from a higher proportion of adverse events (Moorman *et al.*, 2013; Vora *et al.*, 2014). Kaplan-Meier methods, Cox proportional hazards models, and log-rank tests are amongst several techniques employed within this project to determine the impact of varying treatment elements on survival of good risk genetics patients.

1.10 Artificial intelligence

Artificial Intelligence (AI) describes computational programmes that simulate human intelligence, such as problem solving and learning (Shouval *et al.*, 2021). Originally, projects took what is called the knowledge based approach and sought to hard-code knowledge about the world in formal languages, as the computational programmes could then reason automatically about these statements using logical inference rules. However, these projects didn't result in any major success, as people struggle to devise these complex formal rules with enough accuracy to describe the world to the computational programme (Goodfellow, Bengio and Courville, 2016). Thus, machine learning was developed in which the AI systems have the ability to acquire their own knowledge by extracting patterns from raw data (Goodfellow, Bengio and Courville, 2016).

1.10.1 Machine learning

Machine learning (ML) is a branch of computational algorithms that are designed with the intention of emulating human intelligence by learning from the current context with the ability to then apply the algorithm to unseen tasks. A machine learning algorithm is a computational process that uses input data to achieve a desired task without being programmed to produce a particular outcome (El Naqa and Murphy, 2015). These algorithms

automatically adapt through repetition (often called “experience”) so that they can perform the desired task optimally. This process is called training, in which samples of input data are provided along with their desired outcomes. This training can be done continuously through the algorithm’s “life” as it processes new data, rather than a finite process performed during initial adaption. Training typically consists of a training dataset in which the algorithm “learns” and, from its findings, develops a model or identifies factors. These models are then tested in a validation dataset – a previously unseen dataset, in order to test the applicability of the model to other data, determining its efficacy. These two steps can be cycled through multiple times to tune the algorithm. Finally, the optimal model is tested in a third dataset, once again unseen by the algorithm.

There are many tasks that a computational algorithm can perform. The input data can be selected and weighted to provide the optimal outcomes. The algorithm can have a network of possible computational pathways that it arranges for optimal results. It can also determine probability distributions from the input data and use them to predict outcomes (El Naqa and Murphy, 2015). These varied uses make machine learning algorithms widely applicable to a host of different fields including pattern recognition, finance, and entertainment, as well as medical applications.

Machine learning can be classified into 3 types of learning: supervised learning, unsupervised learning, and semi-supervised or “reinforcement” learning (Jordan and Mitchell, 2015). Generally, supervised learning involves building a statistical model for predicting an output based on one or more inputs. A model is trained on data that includes variables (known as features) and the outcome (known as labels). The algorithm then creates a function to map the features to labels, which it can then use to predict the labels of new unlabelled data (Shouval *et al.*, 2021). Unsupervised learning has inputs but no supervising output. This form of machine learning is often used to learn relationships and structure for the data (Shouval *et al.*, 2021; James *et al.*, 2021). The findings from unsupervised learning are then generally evaluated based on performance in a subsequent supervised learning task, assessing whether these findings are useful in the context of the data (Shouval *et al.*, 2021). Reinforcement learning refers to a class of techniques designed to train computational agents to successfully interact with their environment. The learning can happen through trial and error,

demonstration or through a hybrid approach (Esteva *et al.*, 2019). Feedback from the consequences of decisions on the training set shape the model (Shouval *et al.*, 2021).

1.10.1.1 Decision Trees

Decision trees (DT) are supervised learning models that hierarchically map data domains onto response sets (Suthaharan, 2016). Decision trees are utilised in both regression and classification tasks with the goal of predicting a class or value of the target variable where for classification tasks the target variable is discrete, and for regression tasks, it is continuous. This thesis focusses only on the use of classification trees.

Classification decision trees are used to assign a "label" to data such as deciding which class a new observation belongs in. It does this by partitioning the input variables, called features, recursively until a successful classification has been achieved. This classification can be a binary outcome or a categorical variable of three or more, whilst features can be categorical or continuous. A decision tree determines which features it should use to split the tree, as well as the values of those features, at each node by using information gain as a quantitative measure with the aim being to maximise the information gain (Suthaharan, 2016). The decision trees throughout this project use the Gini (generalised inequality index) index as their measure of information gain (defined in Section 2.5.1). The structure of a decision tree is shown in Figure 10.

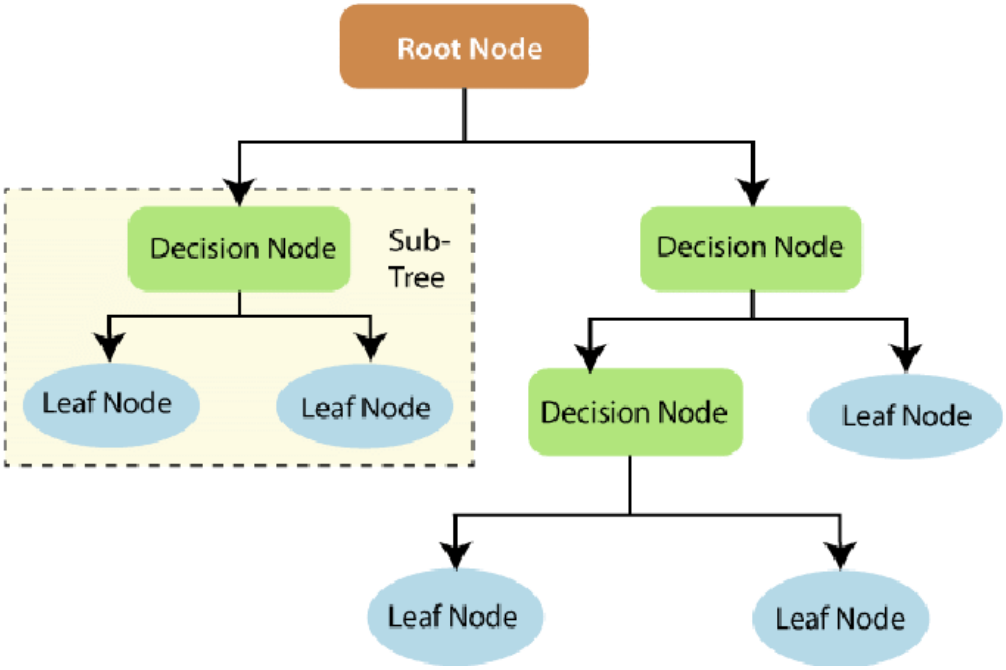


Figure 10. Example structure of a decision tree with the terminology for each part of its construction. (Charbuty and Abdulazeez, 2021).

1.10.1.2 Random Forest

The Random Forest (RF) algorithm is a supervised machine learning method that combines decision trees and aggregates their predictions by voting for classification problems and averaging for regression problems (Biau and Scornet, 2016). Random forest is an ensemble method (Section 2.5.4) as it is an ensemble of trees constructed from training data (Boulesteix *et al.*, 2012). Random forests for classification are focussed on in this thesis.

In the random forest algorithm, an individual training set is acquired for each tree by randomly subsampling overall training set with replacement. This process is referred to as "bootstrap aggregation" or "bagging" (Section 2.5.4.2). The remaining samples from the overall training set are referred to as the out-of-bag (OOB) samples and can be used as internal validation for each individual tree (Rigatti, 2017). Once each tree has been constructed, the predictions of these trees are aggregated through majority voting to attain one overall prediction from the random forest (Parmar, Katariya and Patel, 2019). The performance of the algorithm can then be tested on an external validation set or test set previously unseen by the trees (Section 2.5.3). This process is illustrated in Figure 11.

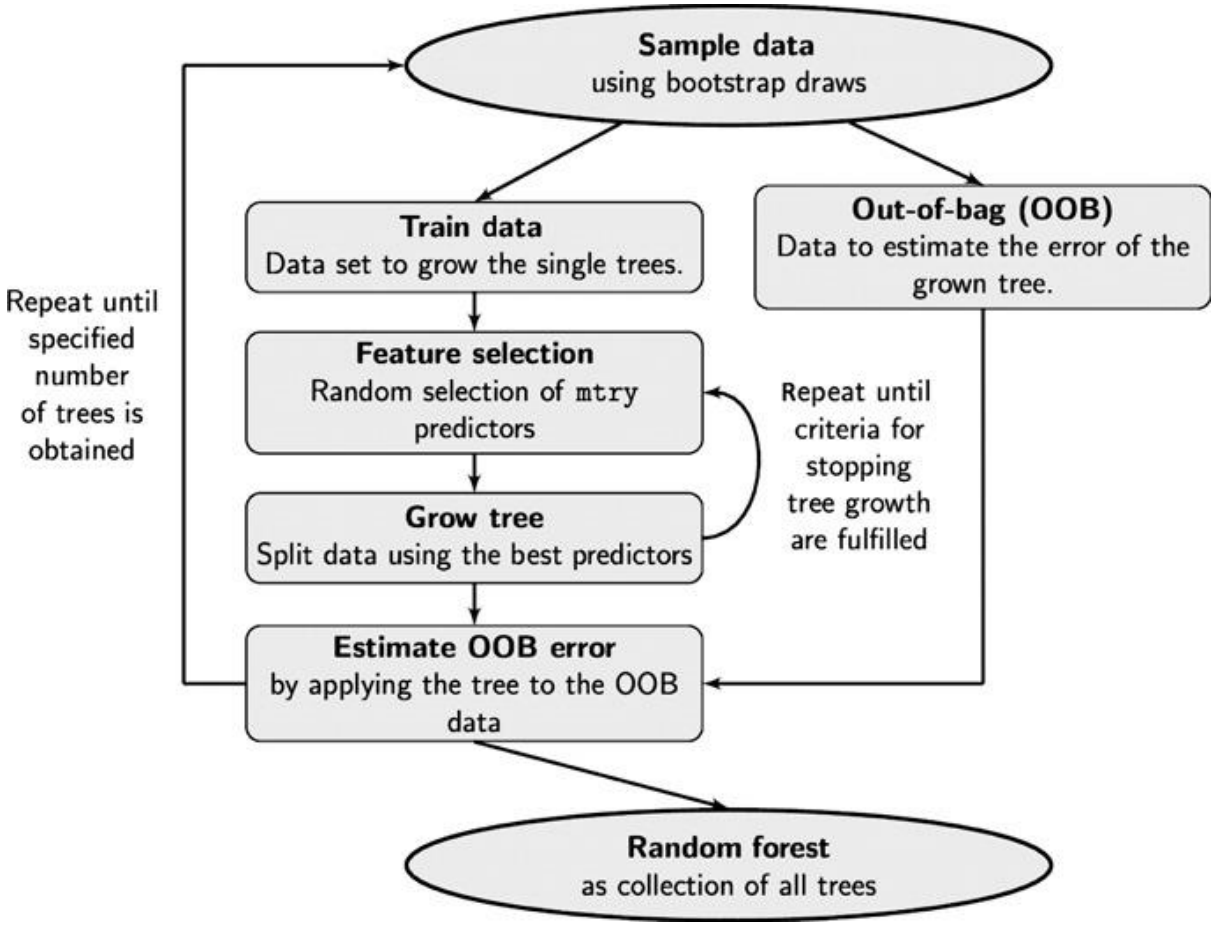


Figure 11. Flowchart demonstrating the process of the Random Forest algorithm. (Boulesteix *et al.*, 2012).

1.10.2 Applications of machine learning

Due to the wealth of data sources in the modern age, ranging from health data such as that of the human genome project, to cybersecurity and social media data, applications of machine learning are vast and continually increasing (Sarker, 2021). Machine learning methods have been applied to various domains including credit risk analysis, fraud detection, medical diagnosis, and customer profiling (Tzanis *et al.*, 2006). Some areas in which machine learning are commonly used, and their specific applications are depicted in Figure 12.

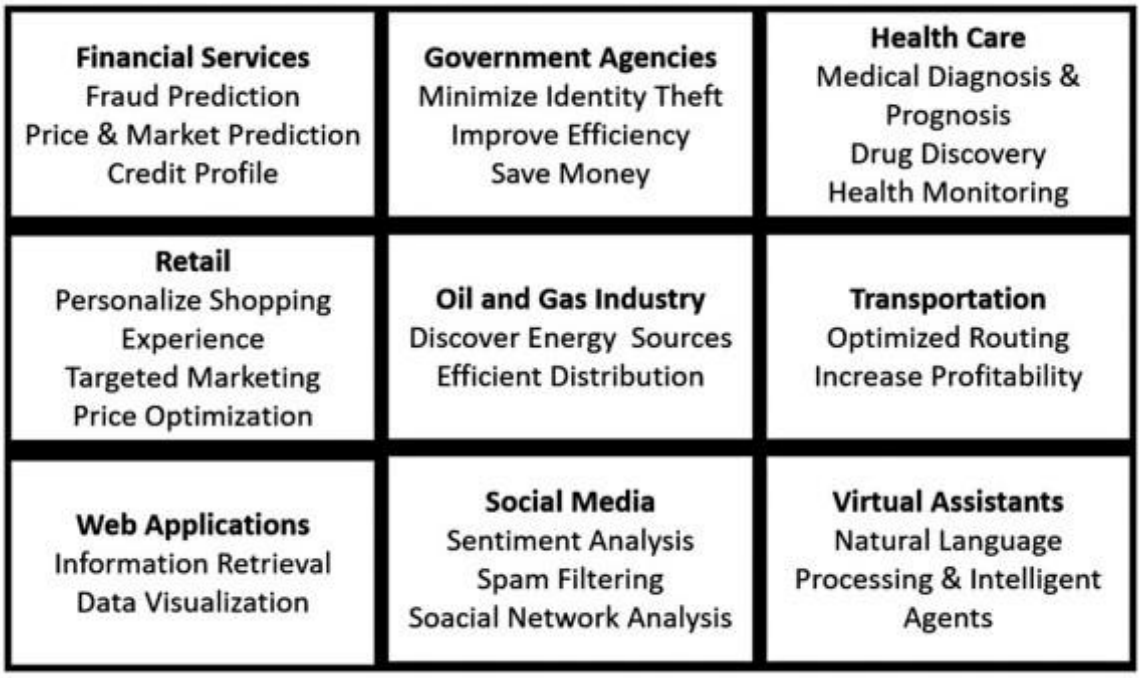


Figure 12. Areas that utilise machine learning. (Geetha and Sendhilkumar, 2023).

In the field of health care, a key area in which machine learning is applied is computer-aided diagnosis. Computer-aided diagnosis uses pattern recognition techniques in order to identify structures of interest from an image such as an X-ray, ultrasound, or MRI (Mohammed, Khan and Bashier, 2016). This utilisation of machine learning is applied within oncology to aid in the diagnosis of several types of cancer included breast, lung, colon, and prostate as well as bone metastases (Mohammed, Khan and Bashier, 2016).

1.10.2.1 Applications in cancer

Machine learning has become increasingly popular in cancer research in a number of areas including cancer susceptibility, relapse prediction and survivability (Kourou *et al.*, 2015). One study showed that the number of papers discussing the use of machine learning in cancer prediction and prognosis increased by 25% per year. It was also found that machine learning has been used to successfully predict outcome or risks in almost a dozen different types of cancer, with the two most prominent being breast and prostate cancer accounting for 24% and 20% of papers in this area respectively (Cruz and Wishart, 2006; Zhang, Shi and Wang, 2023).

Whilst it is clear that there has been much research into the use of machine learning in cancer research, very little has seen to be applied in practice. This could be due to the fact that the field is still relatively new and largely unfamiliar to clinicians and researchers alike. Another theory is that clinicians and oncologists would have to determine the best parameters to be used in predictive models – an area in which they have little or no expertise. Thus, until a model is robust to parameter variation, it will not be useful in practical applications (Park *et al.*, 2013).

1.10.2.2 Applications in leukaemia

Many applications of machine learning in leukaemia are in the interest of diagnosis or classification (Rehman *et al.*, 2018; Shafique and Tehsin, 2018; Rawat *et al.*, 2017; MoradiAmin *et al.*, 2016; Reta *et al.*, 2015; Chin Neoh *et al.*, 2015; Putzu, Caocci and Di Ruberto, 2014; Mohapatra, Patra and Satpathy, 2014; Ongun *et al.*, 2001; Fišer *et al.*, 2012). Table 8 summarises previous applications of machine learning in leukaemia in the literature. Whilst there have been several studies positing the use of machine learning in leukaemia in recent years, machine learning has never been used to address the aim and objectives of this project specifically, highlighting the importance of this work.

Application	Machine Learning Algorithm	Factors included in the model	Efficacy	Reference
Identify children at high risk of ALL relapse	Random Forest Model	14 features including age, WBC, Lymphoblast in BM at diagnosis, Splenomegaly, Hepatomegaly, and BCR::ABL1	Area under the curve (AUC) of 0.902 and 0.904 when applied to two independent test sets	(Pan *et al.*, 2017)
Identify risk factors for adult Philadelphia positive ALL patients	XGBoost (eXtreme Gradient Boosting) (a decision-tree based ensemble ML algorithm)	BCR::ABL1 lineage, age, polymerase chain reaction (PCR), and WBC	Risk groups were defined using the index of dichotomy in the XGBoost decision tree and they were statistically significant in both EFS and OS at 4 years	(Nishiwaki *et al.*, 2021)
Predict HD-MTX-related neutropenia and fever in childhood B-cell ALL	Random Forest Model	Platelets, absolute neutrophil count (ANC), the SNP rs11045879, and delayed clearance of MTX (48hr)	Area under the curve of 0.927 for predicting neutropenia and 0.870 for the model predicting fever	(Zhan *et al.*, 2021)
Predict relapse after an allo-HSCT	Alternating decision tree (ADTree)	Age, diagnosis, the refined disease risk index[56] (rDRI), donor type, graft, the use of TBI, and the conditioning regimen	AUC of 0.667 in the validation cohort, a false positive rate (FPR) of 0.216, a false negative rate (FNR) of 0.5 and an accuracy of 71%	(Armand *et al.*, 2014)
Computer-aided system classifying ALL into subtypes	Convolutional Neural Network (CNN)	Bone marrow images for 4 classifications L1, L2, L3, and normal	Accuracy of 97.78%	(Rehman *et al.*, 2018)
Detection of ALL and classification into subtypes	Deep Convolutional Neural Network (DCNN) – A pretrained AlexNet	Bone marrow images for 4 classifications L1, L2. L3, and normal	Detection: Accuracy of 99.5% Classification: Accuracy of 96.06%	(Shafique and Tehsin, 2018)
Detection of ALL and classification into subtypes	Hybrid Hierarchical Classifiers	Bone marrow images for 4 classifications L1, L2. L3, and normal	Overall accuracy of 97.6%	(Rawat *et al.*, 2017)
Diagnosis of acute lymphoblastic leukaemia	Hierarchical Clustering Analysis (HCA) and Support Vector Machine (SVM)	Samples from bone marrow or peripheral blood	Efficacy not reported	(Fišer *et al.*, 2012)

Table 8. Applications of machine learning in leukaemia within the literature. (Pan *et al.*, 2017; Nishiwaki *et al.*, 2021; Zhan *et al.*, 2021; Armand *et al.*, 2014; Rehman *et al.*, 2018; Shafique and Tehsin, 2018; Rawat *et al.*, 2017; Fišer *et al.*, 2012).

1.11 Project aims and objectives

The overarching hypothesis of this PhD project is that patients with specific genetic abnormalities will respond optimally to different, but specific, treatment combinations. Due to the rise in survival from 50% to ~90% in paediatric ALL, it is reasonable to assume that highly effective chemotherapy regimens for many genetic subtypes exist within historic clinical trial datasets. Currently genetic abnormalities are used within treatment protocols and trials as effective prognostic biomarkers and to drive risk stratification. However, they are rarely used as predictive biomarkers to assign optimal treatment regimens; the exception being the aforementioned *BCR*::*ABL1* and *ABL*-class fusions through adjuvant therapy with a tyrosine kinase inhibitor.

The aim of this project was to identify treatment elements that are optimal for patients with an individual genetic abnormality to ensure that patients are given only the minimal dosages of drugs necessary in a patient's therapy so that they are cured without unnecessary toxicity. The major objectives of the project were as follows:

Objective 1: Compile a single pooled dataset comprising detailed patient-level information across multiple trials. The dataset will comprise three main sections: general patient data, genetic information, and detailed treatment information. The general patient information will include all known risk factors and parameters; for example: sex, age, white cell count, CNS disease, early response information, etc. The genetic section will include standard-of-care diagnostic genetic data collected and reviewed by Profs Moorman and Christine Harrison as part of previous research projects. In addition, we will supplement these data with results from additional genetic screening using FISH, MLPA, SNP arrays, and various NGS techniques using samples collected from patients treated on ALL97/99 and UKALL2003. Detailed treatment data will include which treatment blocks each patient received plus details of randomisations and risk directed therapy (such as cranial radiotherapy).

Objective 2: Utilise standard and advanced statistical techniques to explore the interaction between treatment and genetics. Provide robust data for a handful of genetic subtypes regarding the optimal treatment pathway and use this information prospectively within a clinical trial to assign patients to specific treatments pathways. Determine individual drugs doses for protocol pathways and use this information to derive a treatment intensity index.

Objective 3: Harness machine learning methodologies to explore the genetic-treatment interactions in more depth. Utilise the drug dosages calculated in objective 2 as well as information regarding treatment randomisations, regimens, and intensifications within trials to determine optimal treatment elements for good risk genetic patients.

Chapter 2. Materials and Methods

2.1 Data collection

Data for this project were comprised from multiple paediatric UK trials. Individual patient data for cases treated on UKALLXI, UKALL97, UKALL2003 and UKALL2011 were collected by the Leukaemia Research Cytogenetics Group (LRCG) from the clinical trial unit that conducted the trial and the NHS laboratory that performed the diagnostic genetic testing. Additional genetic data were generated by the LRCG as part of previous research projects. The original clinical trial data files are stored securely by the LRCG in either a bespoke online database or as standalone excel files. The LRCG stores all routine research genetics and genomic data in a bespoke database called CIMS (Cytogenetic Information Management System). All data used within this PhD were extracted directly from these sources and integrated into a single dataset. Both UKALLXI and UKALL97 underwent significant modifications during recruitment which affected outcome so during this PhD the following datasets were considered: UKALLXI, UKALLXI92, UKALL97, UKALL97/99, UKALL2003 and UKALL2011. As screening for *ETV6*::*RUNX1* was not started until UKALLXI92, only the last five datasets were used for this analysis. All patients had given written informed consent for data collection and genetic studies as specified by the trials' protocols.

2.2 Clinical trial data

The data that were utilised in this project were collected from patients enrolled on clinical trials from 1992 – 2018 over the course of four trials.

2.2.1 UKALLXI92 (1992-1997)

Changes to the UKALLXI protocol were made after interim analyses demonstrated the benefit of the early intensification block in 1991. Thus, from March 1992 all patients were to receive two intensification blocks, and a new randomisation was introduced, in which patients received a third intensification block or only two intensification blocks. Daunorubicin in induction was dropped due to concerns about cardiotoxicity. As well as the objectives established for the UKALLXI trial, UKALLXI92 also sought to assess whether a third intensification block improved relapse-free survival and compare long term learning and neuro-psychological effects of different CNS treatments. Randomisation of CNS treatment by presenting WBC remained the same as for UKALLXI, except for children aged between 1 and 2 years at presentation with a white cell count >$50x10^9$/L who were assigned high dose

methotrexate CNS therapy after the modifications to the protocol. An outline of the treatment schedule for UKALLXI92 is shown in Figure 13.

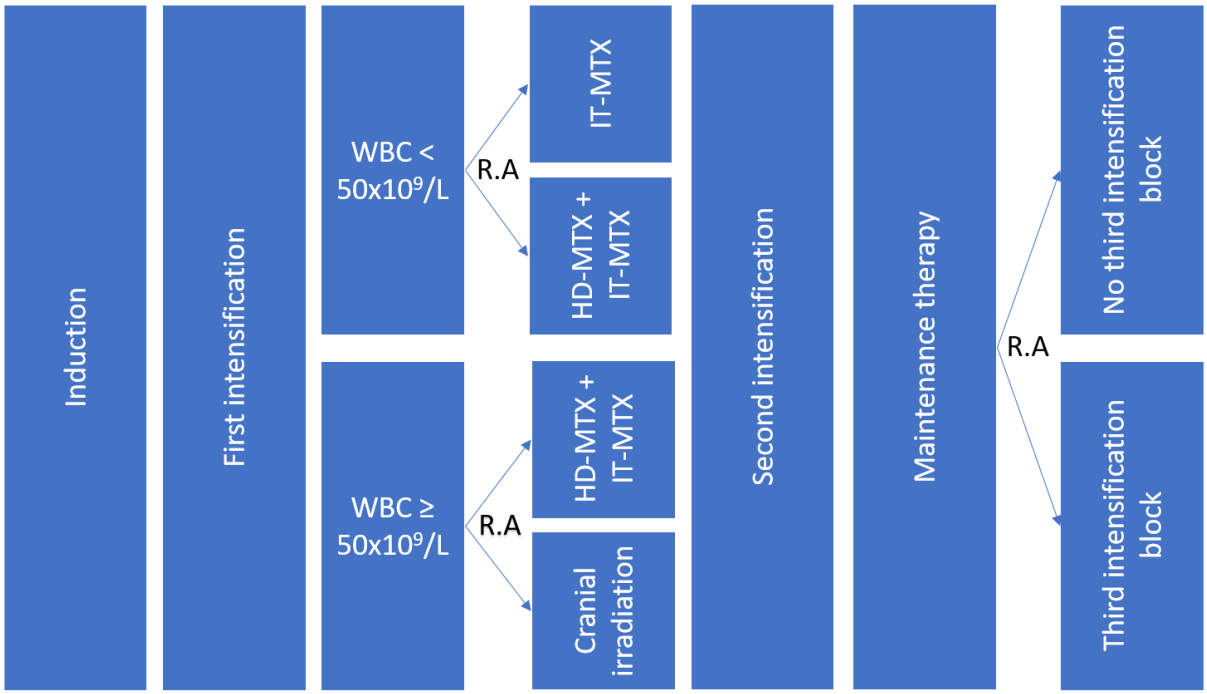


Figure 13. Treatment schedule on UKALLXI92. R.A: Randomly assigned, IT-MTX: Intrathecal methotrexate, HD-MTX: High dose methotrexate, WBC: white blood cell, L: litre.

2.2.2 UKALL97 (1997-1999)

UKALL97 opened in January 1997 to all children aged between 1 and 18 years old except those with mature B-ALL, *BCR*::*ABL1* fusion gene, near haploidy, rearrangements involving the *KMT2A* gene on 11q23 and patients who were characterised as high risk by the Oxford Hazard Score who were treated on HR1 – a high risk study running in parallel to UKALL97. A hazard score of greater than or equal to 0.8 denotes a high score which is calculated as follows: $0.22 \times \log_e (\text{WBC} + 1.0) + 0.0043 \times \text{Age}^2 - 0.39 \times \text{Sex}$ (male = 1; female = 2). In total, 2,108 patients were enrolled on UKALL97 with the objective of determining the effect of steroid and purine randomisations on remission rate and survival as well as continuing the objectives of UKALLXI92 regarding the effect of two vs three blocks of intensive therapy and to assess the need for cranial irradiation. Patients were randomised between the steroids prednisolone and dexamethasone during induction and continuing treatment as well as between the purines mercaptopurine and thioguanine throughout treatment. Further randomisations occurred for

patients with a presenting WBC count > $50x10^9$/L between high dose methotrexate and cranial irradiation for CNS treatment. All patients with a WBC count ≤ $50x10^9$/L were assigned intrathecal methotrexate as their CNS therapy. Continuing from UKALLXI92, patients were also randomised as to receiving a third intensification block or not. Details of these randomisations are outlined in Figure 14.

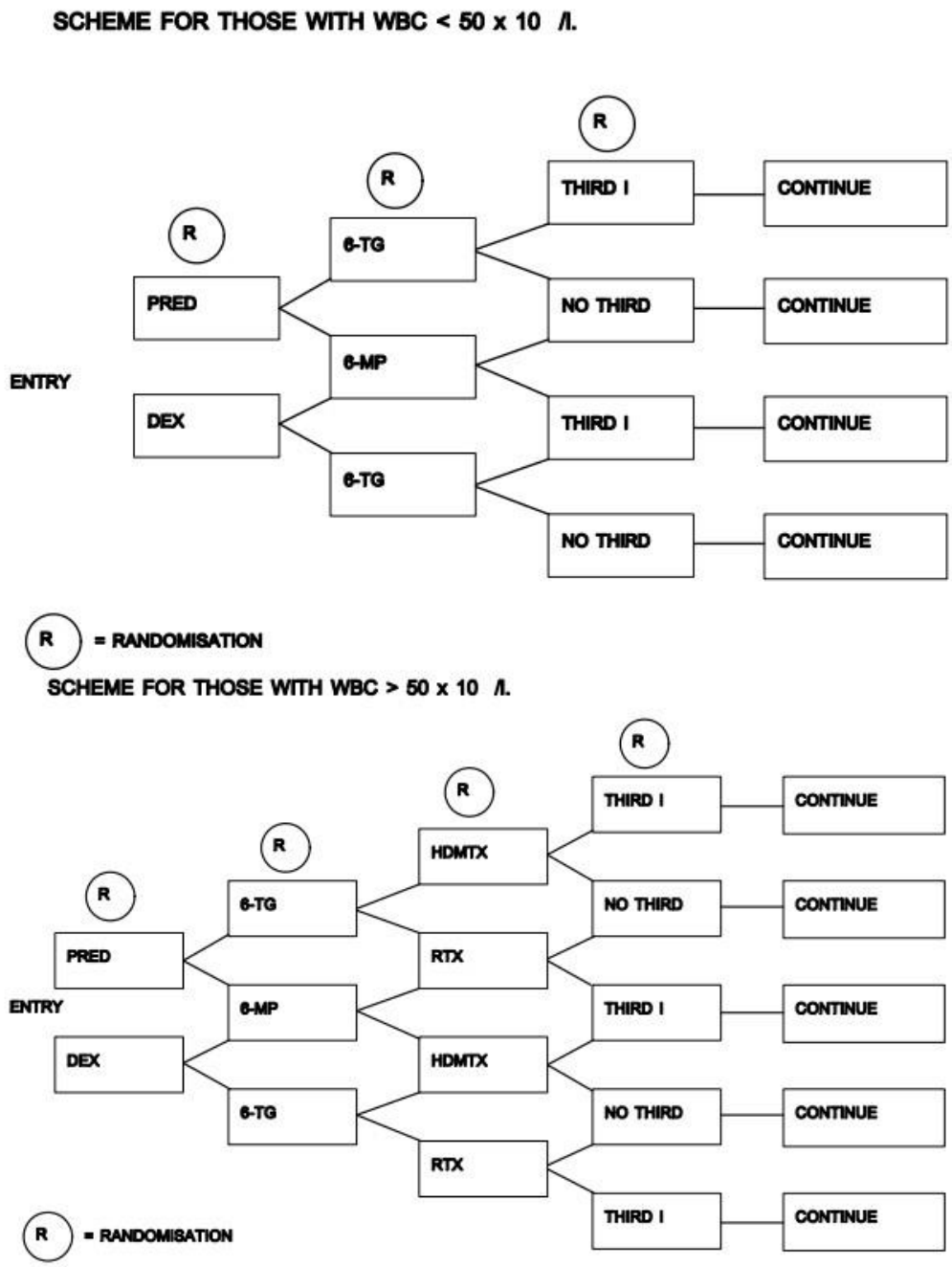


Figure 14. Outline of treatment on UKALL97. WBC: white blood cell, l: litre, pred: prednisolone, dex: dexamethasone, 6-TG: thioguanine, 6-MP: mercaptopurine, HDMTX: high dose methotrexate, RTX: radiotherapy, i: intensification.

2.2.3 UKALL97/99 (1999-2002)

Interim analyses from the UKALL97 trial indicated that there was a significant difference in disease free survival (5-7%) between patients who received a third intensification block and non-recipients, and pharmacokinetic data suggested changing the schedule for asparaginase administration would be beneficial. As such, a modification to the protocol was made in May 1998 in light of these findings. Further protocol changes occurred in November 1999 in which patients were stratified according to NCI risk and intensification modules were modified to comply with the CCG protocol for standard risk patients due to superior event-free survival rates in these patients. On the modified protocol, patients with *BCR*::*ABL1*, near haploidy and *KMT2A* gene rearrangements were now eligible for the trial. Whilst steroid and purine randomisations remained on the modified trial, all other randomisations ceased, with patients being assigned regimens A, B, or C according to their risk stratification [Figure 15]. Patients on regimen C received Capizzi interim maintenance which is a type of maintenance that involves administering escalating doses of methotrexate to a total possible dose of 300mg/m^2 (Viswanathan *et al.*, 2021). Analyses of the trial will be performed in two datasets to account for these modifications – the 1,004 patients enrolled before May 1998 will be referred to as the UKALL97 trial cohort whilst the 1,104 patients enrolled between May 1998 and June 2002 (the end of the trial) will be referred to as the UKALL97/99 trial cohort.

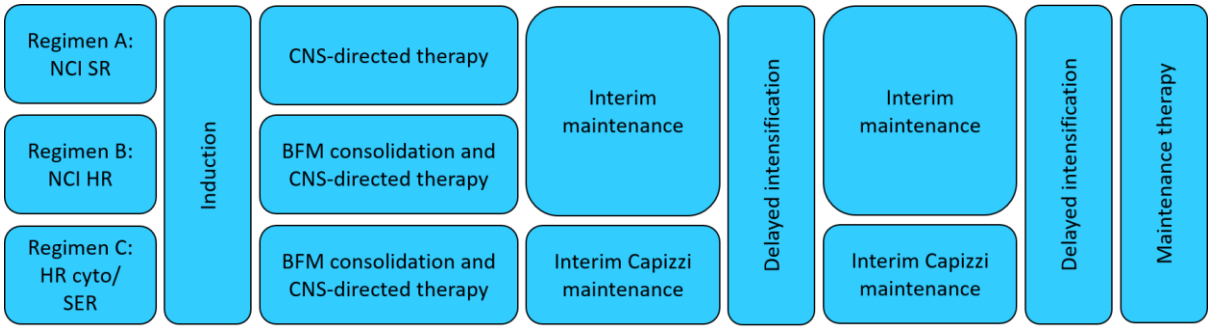


Figure 15. Outline of treatment regimens and randomisations on UKALL97/99. SR: standard risk, HR: high risk, cyto: cytogenetic, SER: slow early response, CNS: central nervous system, BFM: Berlin Frankfurt Münster.

2.2.4 UKALL2003 (2003-2011)

This trial for children and young adults aged between 1 and 25 years old opened in October 2003. All patients with acute lymphoblastic leukaemia were eligible for this trial except those

with a t(9;22) translocation (who were moved onto a different trial after induction) or patients with mature B-ALL. In this trial, patients were no longer randomised for steroids or purines and instead everyone received dexamethasone and mercaptopurine respectively due to findings on UKALL97, which showed an increased risk of CNS and systemic relapses amongst dexamethasone recipients and an increased risk of Veno-Occlusive Disease (VOD) of the liver in thioguanine recipients.

A total of 3,207 patients were enrolled on this trial and were initially split into three risk groups which were used to assign their treatment regimen. These groups were standard, intermediate, and high risk, which were used to assign patients to regimen A and regimen B or regimen C respectively. Standard risk patients were defined as children < 10 years old with an initial white cell count $<50x10^9$/L who do not have *BCR*::*ABL1*, hypodiploidy (≤44 chromosomes), or a *KMT2A* gene rearrangement. Intermediate risk was for patients ≥ 10 years old, and/ or children with a diagnostic white cell count ≥ $50x10^9$/L without *BCR*::*ABL1*, hypodiploidy (≤44 chromosomes), or a *KMT2A* gene rearrangement. Patients were assigned to high risk if they had certain genetic abnormalities: a *BCR*::*ABL1* rearrangement (induction only), low hypodiploidy, near haploidy, a *KMT2A* gene rearrangement, or iAMP21. Children who had a slow early response, defined as having more than 25% blasts in the marrow transferred to regimen C from their initial regimen along with the patients classified as high risk. This was at day 15 for regimen A patients, and for regimen B cases this was at day 8 unless they were ≥16, in which case they remained on regimen B. At day 28 patients on regimen A and B has their bone marrow status reassessed and if they had >5% but <25% blasts they were transferred to regimen C. If their status was >25% blasts they were taken off protocol. If patients on regimen A and B had <5% blasts at day 28, they remained on their initial regimen.

Children on UKALL2003 received either 1 or 2 delayed intensifications as outlined below. Patients were defined as MRD high risk if they had an MRD level $>10^{-4}$ at day 29 and MRD low risk if they had negative MRD or an MRD level $<10^{-4}$ at day 29. High risk patients were randomised for increased intensity: remaining on regimen A or B vs transferring to regimen C with 2 delayed intensifications. Low risk patients had their MRD levels repeated at week 11 and if it was negative at week 11 they were randomised between 1 and 2 delayed intensifications and if it was positive then patients were assigned 2 delayed intensifications.

Patients with indeterminate MRD at either time point as well as all patients on regimen C, regardless of MRD status, received 2 delayed intensifications. Minor changes to the protocol were made throughout the trial including a change to MRD randomisations in August of 2009. From this point, all patients who achieved MRD low risk status received 1 delayed intensification. The threshold for low risk status also increased to $<5 \times 10^{-5}$ and MRD low risk patients no longer needed the repeat level at week 11 and were instead categorised at day 29. The protocol for MRD high risk patients and those with indeterminate MRD remained the same. An outline of these details is given in Figure 16.

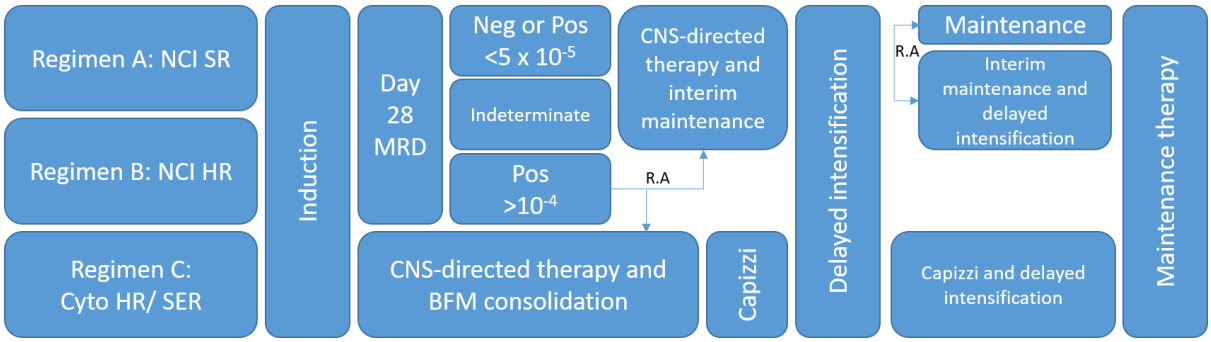


Figure 16. Outline of treatment regimens and randomisations on UKALL2003. SR: standard risk, HR: high risk, SER: slow early response, R.A: Randomly assigned, neg: negative, pos: positive, cyto: cytogenetic, MRD: measurable residual disease, CNS: central nervous system, BFM: Berlin Frankfurt Münster.

2.2.5 UKALL2011 (2011-2018)

2,750 patients were registered on the UKALL2011 trial between April 2012 and December 2018 when the trial closed. The aim of this study was to define whether refinement of MRD based risk stratification and treatment regimens improved survival whilst reducing overall burden of therapy in patients. There were also randomised and non-randomised objectives. The randomised objectives were: (1) to reduce toxicity through induction of a short 14 day course of high dose dexamethasone in lieu of the conventional lower dose given for 28 days and (2) To provide more effective CNS prophylaxis and reduce burden of therapy through introduction of high dose methotrexate, and by omission of vincristine and dexamethasone pulses and continuing intrathecal therapy in maintenance. The non-randomised objective was to decrease toxicity and reduce burden of therapy by administering a single delayed

intensification to all patients and limiting augmented therapy to those who are not MRD low risk.

On UKALL2011, patients who had a slow early response did not move to regimen C until the start of consolidation and instead remained on their original regimen for induction. This was also the case for patients on regimen B with high risk cytogenetics. However, patients on regimen A with high risk genetics moved to regimen C on day 15. Furthermore, *BCR*::*ABL1* patients were immediately transferred to a different trial/ protocol.

All patients received 1 delayed intensification on this trial. There were 3 randomisations on this trial: (1) In induction, patients were randomised between short vs standard dexamethasone in which the drug is scheduled at 10mg/m^2/day for 14 days with no taper or the standard schedule of 6mg/m^2/day for 28 days with a taper. (2) The second randomisation occurs in interim maintenance where patients receive standard interim maintenance where seven weekly doses of 20mg/m^2 of oral methotrexate are scheduled for patients, or high dose methotrexate (protocol M) in which patients receive 5g/m^2 four times over the 9-week phase. (3) Finally, in maintenance therapy, patients are randomised between receiving pulses of vincristine and dexamethasone on top of the maintenance backbone of mercaptopurine and methotrexate, and not receiving the pulses. A visual representation of the treatment regimens and randomisations is shown in Figure 17.

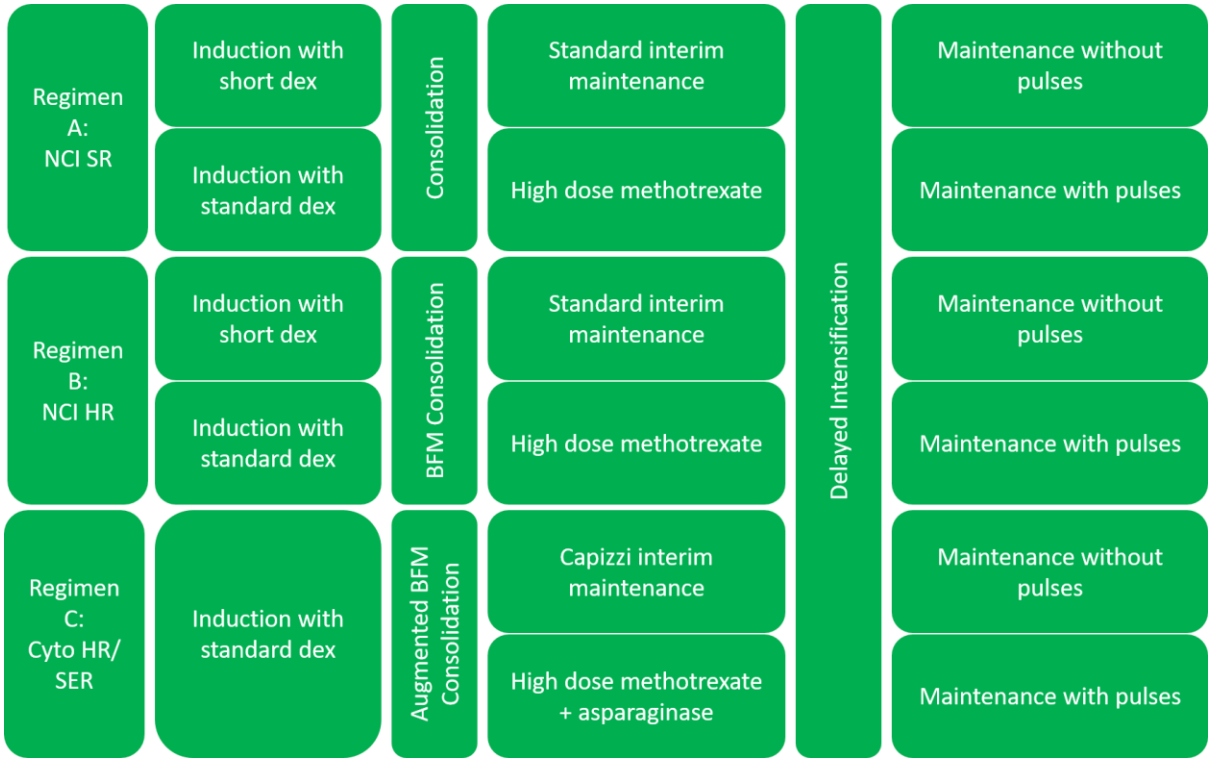


Figure 17. Outline of treatment regimens and randomisations on UKALL2011. SR: standard risk, HR: high risk, cyto: cytogenetic, SER: slow early response, dex: dexamethasone, BFM: Berlin Frankfurt Münster.

2.3 Data processing

After exporting all the relevant variables for this project from the aforementioned sources, there was a large volume of data as demonstrated in Table 9. Due to this, three datasets were created for each trial, except UKALL2003 and UKALL2011 for which suitable datasets were already available. The variables were split in demographic, genetic, and treatment datasets respectively. After cleaning and formatting, these datasets were then merged to form a final dataset containing any pertinent data.

Number of Patients	Number of Variables
UKALLXI92	
1709	197
UKALL97	
1004	235
UKALL97/99	
1104	197
UKALL2003	
3112	158
UKALL2011	
2517	183

Table 9. Number of variables and patients available for each trial in the project.

2.3.1 Data formatting

The process of formatting the data began by ensuring that each of the categorical demographic variables such as sex and CNS status were identically coded across the trials, (e.g. 0 = male and 1 = female rather than 1 = male and 2 = female) in order to avoid errors when merging the datasets to create the final dataset used for analysis. Any variables that held the same information or any individual variable with multiple entries that held the same information were compiled so that it was all available only once in the dataset. This was a particularly common occurrence for genetic variables as they are often named in multiple ways to refer to the same abnormality (eg. Philadelphia chromosome = t(9;22) = *BCR*::*ABL1*). Further to this, dates were all assigned the same format of DD/MM/YYYY and the data were inspected, confirming that the dates occurred in a chronological order (eg. date of birth, date of treatment start, date of relapse, date of death). Finally, the date a person was last seen was changed to the latest date in the dataset using the "max" command as this was necessary to make certain any survival rates calculated from these variables were accurate.

2.3.2 Data cleaning

In order to create an optimal final dataset for analyses, any variables that were not pertinent to the analysis, as well as any duplicate variables were removed from the dataset. Patients who were no longer eligible for analyses due to misdiagnoses, withdrawal of consent, or

withdrawal from the trial were also removed. Files containing information relevant to these data held in LRCG stores were studied to identify potential protocol deviations and amend these data or remove these patients from the analysis accordingly. Examples of this include patients who moved treatment regimen due to patient request or patients who did not receive the allocated regimen but there is no recorded information on which treatment arm they did receive. This resulted in a finalised dataset of 6,716 patients who were eligible for analyses from the four trials and 5,880 patients with the necessary information available to calculate drug dosage. These patients ranged in age from <1 – 24 years old with a mean age of 6 years old. A consort diagram outlining the method leading to the finalised dataset is shown in Figure 18.

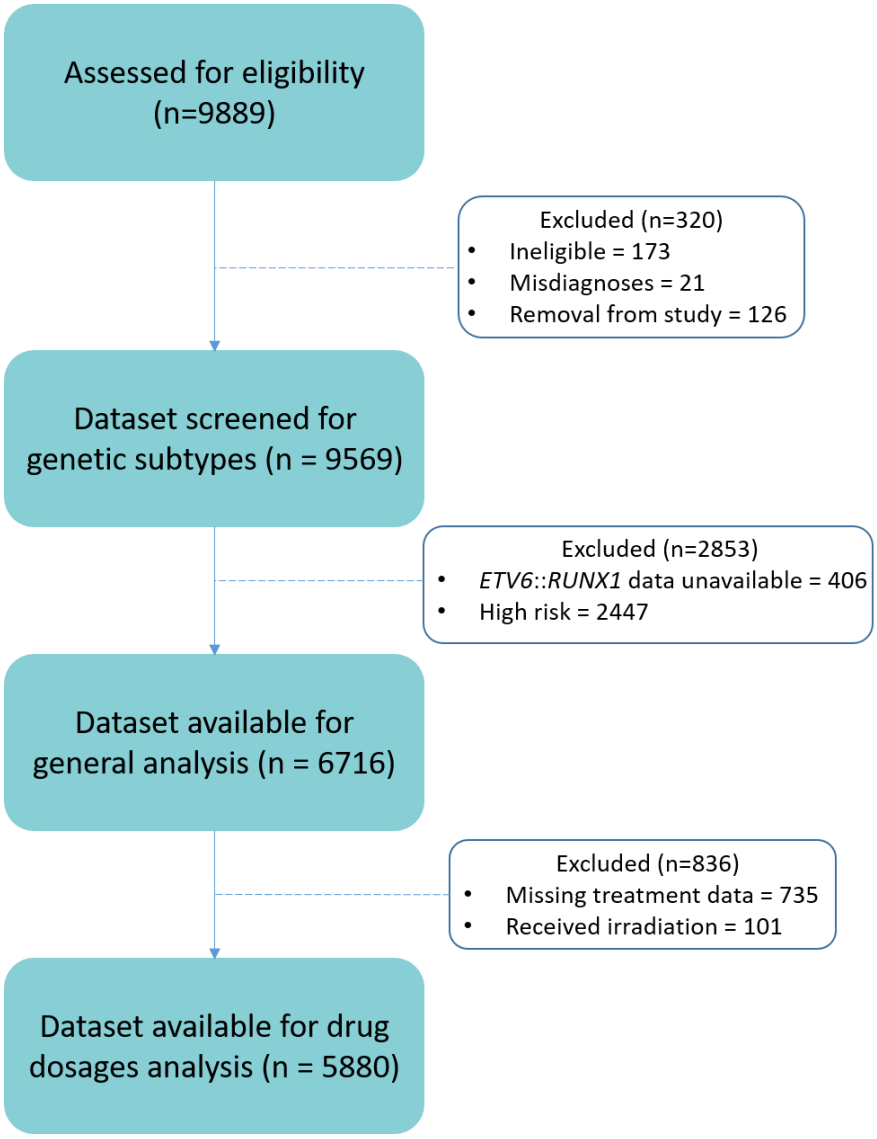


Figure 18. Consort Diagram of method leading to the finalised dataset.

2.4 Statistical analysis

Two endpoints were considered in this project: overall survival (OS) and event-free survival (EFS). Overall survival was defined as time from registration onto trial to death, censoring at date of last contact. Event-free survival was defined as time from registration to relapse, second tumour, or death, censoring at date of last contact. All survival rates were estimated using Kaplan-Meier methods and are quoted at 5 years and plotted using Kaplan-Meier curves. Risk tables were included in place of censor points on the Kaplan-Meiers to ensure clarity of curves whilst still retaining all necessary information regarding censored patients. The two-sided log rank test was used to evaluate the equality of survivorship functions in different subgroups. Univariate Cox proportional hazards models were used to investigate predictors of OS and EFS. The proportional hazards assumption states that the relative hazard must remain constant over time with different predictor or covariate levels. This was assessed using 'estat phtest' function in STATA. Mann-Whitney U tests were used to compare medians and assess distributions across continuous variables and comparison of categorical variables were performed with X^2 test or Fisher's exact test as appropriate. Forest plots and tests of heterogeneity, namely Cochran's Q, I^2, H^2, and τ^2, were used to compare hazard ratios across different subgroups. P-values < 0.05 were considered statistically significant. Analyses were assumed to have adequate power due to a large cohort size (9163 patients total) surpassing any number of patients needed for 90% power in a log-rank test based on the Freedman method (Freedman, 1982). All analyses were performed using Intercooled Stata 18.0 (StataCorp, College Station TX) or R 4.3.2 (R Core Team, Vienna Austria).

2.4.1 T-stochastic Neighbour Embedding (t-SNE) clustering

T-stochastic Neighbour Embedding is a technique to visualise high-dimensional data by converting the data into a matrix of pair-wise similarities (Van der Maaten and Hinton, 2008). It is most commonly used to produce a two-dimensional embedding of data in order to simplify the identification of structures in the data such as clusters (Linderman and Steinerberger, 2019; Arora, Hu and Kothari, 2018). Unsupervised t-SNE was used to visualise drug dosage and randomisation data to assess the presence of clusters based on patient outcome. Analysis was carried out using the "TSNE" function in Python's scikit-learn package (version 1.4.2), Values for perplexity were assessed and adjusted dependent on the datasets

analysed. For all other parameters, default settings were used. Figure 19 exemplifies a t-SNE plot.

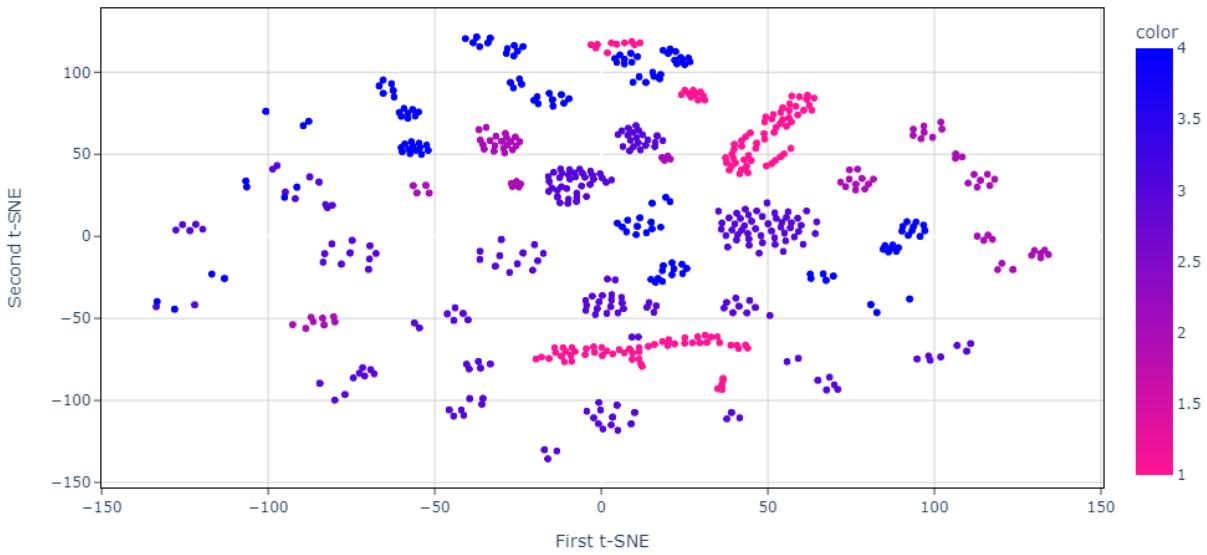


Figure 19. An exemplar t-SNE plot.

2.5 Machine Learning

Decision tree, Random Forest, and XGBoost algorithms were employed in this project to determine optimal treatment elements to classify patients by outcome. These models were optimised (Section 2.5.2), and assessed using standard metrics (Section 2.5.3).

2.5.1 Gini Index

The Gini index was used to assess the purity of each node. A dataset is considered pure if all the samples belong to the same class. The Gini index is a measure of how mixed the dataset is at each node. The range for the Gini index is between 0 and 0.5 where 0 is a completely pure dataset and 0.5 represents a completely impure dataset (Pramod, 2023). The Gini index is defined as:

$$\text{Gini} = 1 - \sum_i p_i^2,$$

where p_i is the probability of an object being classified to a particular class (Thakar, 2022).

2.5.2 Optimisation

Decision tree algorithms are considered "greedy" as they consider the best split at each individual node without considering any previous split and without looking back after a split,

i.e. all of the features are considered for split at each node regardless of if they have been used before and no pruning technique is applied (Zharmagambetov *et al.*, 2021). Whilst this process often results in highly accurate trees that classify the training data well, these trees often don't generalise well to unseen data due to over-fitting or over-parameterising the problem (Bennett, 1995). As such, optimisation methods are often applied to decision trees to ensure their generalisability. Optimisation methods applied to the machine learning methods within this thesis are described below and include cross-validation and pruning.

2.5.2.1 Cross-Validation

Cross-validation has a wide range of applicability in machine learning including tasks such as accuracy estimation, feature selection or parameter tuning (Blockeel *et al.*, 2003). K-fold Cross-validation to estimate the predictive accuracy of the decision tree is performed by building K trees from subsets of the full dataset and assessing their accuracies to ensure it remains stable across these subsets (Galathiya, Ganatra and Bhensdadia, 2012). This ensures the classifications of our decision tree are truly accurate and not simply a result of bias in the training data (Nti, Nyarko-Boateng and Aning, 2021).

2.5.2.1.1 Leave-one-out cross-validation

Leave-one-out (LOO) cross-validation is a type of K-fold cross validation in which K equals n-1 observations in a dataset of size n (Wong, 2015; Magnusson *et al.*, 2020). The model is trained on all the observations except one, then the outcome of this observation is predicted. This is then repeated for all n observations (Magnusson *et al.*, 2020). This is particularly advantageous when the sample size is small as it ensures a maximum number of samples in the training dataset, giving the model the best possible chance of correctly classifying patients due to more examples.

2.5.2.2 Pruning

Pruning is a technique employed to both reduce overfitting of a decision tree and to simplify the tree to increase the interpretability (Mohamed, Salleh and Omar, 2012). This is done by reducing the size or depth of a decision tree by removing sections from the tree so as not to compromise the algorithms ability to accurately classify samples (Shamrat *et al.*, 2021). Pruning can be performed in two different ways. The first is to prospectively decide when to stop the growth of the tree, a method that is referred to a pre-pruning; the second is to

retrospectively reduce the size of the full tree, referred to as post-pruning (Esposito *et al.*, 1997). There are several established methods for each of these approaches two of which will be utilised in this project.

2.5.2.2.1 Cost Complexity

Cost-complexity pruning is a post pruning technique comprised of two steps. Where T_{max} is the largest possible unpruned tree, the following is performed:

1. Selection of a parametric family of subtrees of T_{max} obtained by sequentially pruning branches from T_{max} that results in the lowest increase in apparent error rate per pruned leaf.
2. Choice of the best tree T_i according to a true error rate with respect to the predictive accuracy of the trees in the parametric family (Esposito *et al.*, 1997).

The optimal choice for a tree is one that maximises accuracy with the fewest number of nodes in order to avoid overfitting or underfitting the tree, both of which would lead to a tree with poor generalisability.

2.5.2.2.2 GridSearchCV

GridSearchCV is a function in python's scikit-learn package (version 1.4.2) that can be used to perform hyperparameter tuning. Hyperparameters are second-level tuning parameters usually specified by the designer of the model, and include elements such as the maximum depth of the tree or the minimum number of samples per leaf; as opposed to the first-level model parameters which are learned by the algorithm from the training data (Probst, Boulesteix and Bischl, 2019; Nyuytiymbiy, 2020). Hyperparameter tuning is the method of comparing a range of values of a model's hyperparameters to choose the values that optimise the model on certain metrics such as accuracy (Probst, Wright and Boulesteix, 2019). GridSearchCV iterates through all the possible combinations of values for the hyperparameters in a model to choose both the best hyperparameters for a model and their optimal values (Ahmad *et al.*, 2022). This can be used as a pruning technique for decision trees as GridSearchCV can iterate through possible depths of a tree, as well as the number of features to consider at each node, resulting in an optimal max depth and max number of features for a highly accurate tree.

2.5.2.3 Addressing imbalanced classes

A common problem that can affect the efficacy of a decision tree algorithm is that of imbalanced classes (He and Garcia, 2009). Data are imbalanced if the classification categories are not approximately equally represented (Liu, Wang and Zhang, 2009). This is an issue as several machine learning algorithms, including decision trees, assume approximately balanced classes (Tanha *et al.*, 2020). This imbalance can result in high global accuracy for a decision tree as the majority class can be correctly classified, but may have incredibly low accuracy in the minority class, which is particularly troublesome when this is the class of interest (Vuttipittayamongkol, Elyan and Petrovski, 2021). There are several widely accepted approaches to resolve the issue of imbalanced data and thus improve the accuracy in the minority classes, many of which include data resampling (Patel *et al.*, 2020). Several methods employed in this thesis include oversampling, undersampling, and weighting which are outlined in detail below. The python package imbalanced-learn (0.12.3) was used for SMOTE and NearMiss resampling whilst the weighting was embedded in the DecisionTreeClassifier command within scikit-learn (1.4.2).

2.5.2.3.1 Oversampling

Oversampling is an approach that balances the data by adding additional samples to the minority class (Krawczyk, 2016). These samples may be duplicated from the minority class or synthetic data created through techniques utilising nearest neighbour strategies (Fernández *et al.*, 2018b). A visual demonstration of this is shown in Figure 20. The oversampling method used in this project is the Synthetic Minority Over-Sampling Technique (SMOTE) which follows this process:

1. Take a sample from the minority data and consider its k nearest neighbours in the feature space
2. Take the vector between one of those k neighbours and the minority sample.
3. Multiply this vector by some number, x, which lies between 0 and 1.
4. Add this to the minority sample to create the synthetic data point (Chawla *et al.*, 2002).

Figure 20. Examples of resampling methods. (Mohammed, Rawashdeh and Abdullah, 2020).

2.5.2.3.2 Undersampling

Undersampling is resampling approach that balances the data by excluding majority samples from the dataset (Krawczyk, 2016). These samples may be selected randomly or through nearest neighbour methods such as the NearMiss algorithm (Fernández *et al.*, 2018b). There are three versions of the NearMiss algorithm: NearMiss-1, which selects majority samples by determining those whose average distance to the three closest minority samples is the smallest. NearMiss-2 selects majority samples whose average distance to the three farthest minority class samples is the smallest. NearMiss-3 selects a given number, n, of the closest majority samples to each minority sample to ensure that every minority sample is surrounded by some majority samples. NearMiss-1 was used in this project.

2.5.2.3.3 Weighting

A technique to address imbalanced data without resampling is that of class weights. This is implemented by assigning weights to samples in the different classes, which informs the model how much "attention" to pay to these samples thus stopping the model from ignoring samples in the minority class (Fernández *et al.*, 2018b). Some common methods include assigning equal weight to the classes so that the model assigns equal importance to correctly classifying the groups, or assigning a higher weight to the minority data and a lower weight to the majority data in order to have the model prioritise correctly classifying minority samples (Huang *et al.*, 2013; Vuttipittayamongkol, Elyan and Petrovski, 2021).

2.5.3 Model Evaluation

It is vital that the decision tree model is evaluated to ensure the classifications are accurate. This is performed by employing the decision tree on unseen data, often referred to as test data, and measuring its performance. There are several metrics to assess the performance of decision trees on unseen data many of which involve the use of the true positive, false positive, true negative and false negative rates as detailed in Table 10. In python, the below metrics were determined using commands from the scikit-learn (1.4.2) package.

Ratio	**Formula**	**Definition**
True positive rate (Sensitivity)	$\frac{TP}{TP + FN}$	The probability that an actual (true) positive result will test positive.
False positive rate	$\frac{FP}{FP + TN}$	The probability that a positive result will be given when the actual (true) value is negative.
True negative rate (Specificity)	$\frac{TN}{TN + FP}$	The probability than an actual (true) negative result will test negative.
False negative rate	$\frac{FN}{FN + TP}$	The probability that a negative result will be given when the actual (true) value is positive.

Table 10. Ratios used in measuring the accuracy of a decision tree.

2.5.3.1 Accuracy

A common metric to determine the overall performance of a decision tree is the accuracy score. The accuracy score is used to measure the proportion of correctly classified instances in the test data. Accuracy scores are often denoted as a decimal and range from 0 to 1 or can be viewed as a percentage. The formula for the accuracy score is:

$$Accuracy\ score = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP = true positive, TN = true negative, FP = false positive and FN = false negative. This can be interpreted as the total number of correct predictions over the total number of predictions made (Vujović, 2021).

Whilst the accuracy score is a useful metric, in certain instances a decision tree can have a high accuracy score and still be a poor model overall due to issues such as imbalanced classes (Section 2.5.2.3). In these instances, other, more appropriate metrics can be considered such as precision, recall, and F1-score.

2.5.3.2 Precision

Precision, also called specificity, is a ratio of the true positive cases out of all cases who were predicted to be positive. This metric is a decimal that ranges between 0 and 1 where a higher value denotes a larger proportion of accurate positive predictions. Precision can be written mathematically as:

$$Precision = \frac{TP}{TP + FP}$$

Where TP is true positive and FP is false positive (Vujović, 2021).

This metric is a useful aid in determining the overall performance of a decision tree as it ensures that the tree doesn't incorrectly classify many patients as positive, which, in the instance of imbalanced classes, could still result a high accuracy score. Precision can be interpreted as how often the decision tree is correct when it has given a positive prediction.

2.5.3.3 Recall

Recall, also called sensitivity, is the measure of how frequently the decision tree has correctly classified the true positive cases. The formula for this is:

$$Recall = \frac{TP}{TP + FN}$$

Where TP is true positive and FN is false negative (Vujović, 2021).

This metric is a measure of how many cases of an event of interest have been identified. Whilst the overall aim of a decision tree is perfect classification with no errors, in practice this

is highly unlikely. When a decision tree doesn't have a perfect classification, it isn't possible to have high precision and recall simultaneously, as one is increased at the cost of the other. This is referred to as the precision/recall trade-off (Kuruvilla and Kundapura, 2022).

2.5.3.4 F1-Score

F1-score is a useful metric in instances where precision and recall are equally important and a trade-off isn't possible as it is the harmonic mean of the precision and recall (Kuruvilla and Kundapura, 2022). It can be written mathematically as:

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

F1-score can also be depicted as a decimal or a percentage and a high F1-score indicates high precision and recall values (Vujović, 2021).

2.5.3.5 Confusion Matrix

A confusion matrix depicts the results of the classification performed by the machine learning algorithm (Varoquaux and Colliot, 2023). It divides the samples into four categories (for binary classification) based on their true labels and predicted labels demonstrating how many were accurately classified or misclassified (Varoquaux and Colliot, 2023). An example confusion matrix is shown in Figure 21 below.

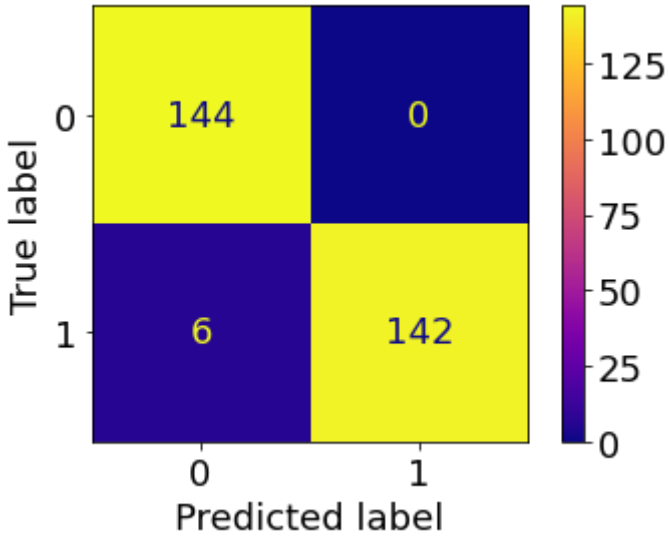


Figure 21. Example confusion matrix.

2.5.3.6 Receiver Operating Characteristic Curve

A receiver operating characteristic (ROC) curve is the plot between the sensitivity (y-axis) and 1-specificity (x-axis) which shows the performance of the decision tree at all classification thresholds (Vujović, 2021; Nahm, 2022). As seen in Figure 22, the baseline of an ROC curve is the straight line from the point (0,0) to (1,1) on the plot and this would indicate a decision tree with random performance. Any curve above the baseline is considered a good indicator of the trees performance whilst anything below the baseline suggests poor performance (Nahm, 2022).

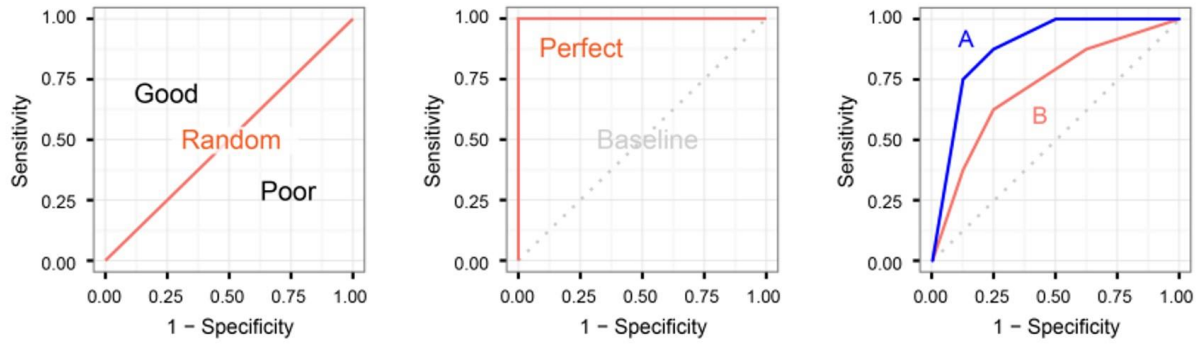


Figure 22. Examples of Receiver operating characteristic curves indicated performance. (Saito, 2024).

A common measure used when evaluating the performance of a decision tree using the ROC curve is the area under the ROC curve (AUC) (Muschelli, 2020). This is the area under the curve calculated in the ROC space and has a range between 0 and 1 where the midpoint, 0.5, is the score of a random classifier, the baseline curve (Hoo, Candlish and Teare, 2017). AUC scores of greater than 0.7 are considered good classifiers with values of 0.8 and above generally preferred for a model with practical applications (Hosmer Jr, Lemeshow and Sturdivant, 2013).

2.5.4 Ensemble methods

An approach used to improve the accuracy and performance of individual machine learning algorithms is the use of ensemble methods (Mohammed and Kora, 2023). In the context of classification, ensemble methods are learning algorithms that construct multiple individual or "base" learners and combine them, often by aggregating the classification predictions of the learners and taking the majority vote (Dietterich, 2000; Kunapuli, 2023). These base learners

are usually machine learning algorithms such as decision trees or neural networks (Zhou, 2012). As a result of this process, ensemble methods are often better at generalising (accurately predicting classification of unseen data) than the base learners and are thus frequently referred to as strong learners. Conversely, base learners are referred to as weak learners (Zhou, 2012). As well as the original ensemble method of Bayesian averaging, there are also Boosting (Schapire, 1990; Freund, 1995; Freund and Schapire, 1996) and Bagging (Breiman, 1996) approaches to ensemble learning.

2.5.4.1 Boosting

The concept of boosting is to produce a series of classifiers where the training set used for each classifier is based off of the performance of the earlier classifiers in the series, resulting in incorrectly classified samples being selected more often than correctly classified samples (Maclin and Opitz, 1997). This gives the algorithm more opportunity to correctly classify those samples, whilst ignoring the ones already correctly classified. Extreme Gradient Boosting (XGBoost) was the boosting algorithm applied in this project.

2.5.4.2 Bagging

Bagging is a term derived from a technique called bootstrap aggregation and is a parallel learning approach (Kumar, Kaur and Gosain, 2022). This ensemble method is performed by sampling, with replacement, the instances in the training dataset to produce a training dataset for each base learner (Maclin and Opitz, 1997). As sampling is done with replacement and the probability of selecting a sample is equal across the training set, samples are likely to appear in multiple training sets (Sagi and Rokach, 2018; Maclin and Opitz, 1997). Similarly to Boosting, a final classification is determined by majority voting of the classifications of the weak learners (Kumar, Kaur and Gosain, 2022). However, unlike Boosting, equal weights are assigned to each learner and thus they all have equal influence on the final vote. Random Forest was employed as a bagging technique in this project.

2.5.5 Representation learning

The performance of machine learning algorithms depends on the representation of the data they are given (e.g. pictures of apples and oranges to teach an algorithm to distinguish between the two). The importance of representations can also be found in other contexts – for example, humans can easily perform arithmetic on Arabic numerals but find arithmetic on

Roman numerals more difficult. It is important to identify the optimal set of features for the algorithm to perform the task effectively, however, this is difficult in practice (Goodfellow, Bengio and Courville, 2016). Representation learning offers a solution to this problem by allowing the machine to be fed with raw data and to discover the representations needed for classification. Deep learning methods are representation learning methods with multiple levels of representation in which the representation at one level is transformed by a non-linear model into a representation at a higher level (LeCun, Bengio and Hinton, 2015).

Chapter 3. Utilisation of survival analysis methods to identify optimal treatment elements for cure of patients with good risk genetics

3.1 Introduction

The good risk genetic subgroups account for ~50% of childhood ALL cases and are associated with a favourable prognosis, with survival rates >90% at five years for both *ETV6*::*RUNX1* and high hyperdiploidy patients. The overwhelming majority of these patients are enriched for good prognostic features, having low levels of minimal residual disease, low white cell counts, and being below the age of 10 years old (Bhojwani *et al.*, 2012; Paulsson and Johansson, 2009). Due to these factors, *ETV6*::*RUNX1* and high hyperdiploidy patients are most often treated on lower intensity protocols. Many studies have sought to further refine stratification of these good risk genetic subgroups based on demographic and genetic features, with the goal of reducing treatment intensity in subsets of patients with excellent prognosis, or identifying patients with a poorer outcome. Enshaei *et al.* identified a low-risk subgroup within high hyperdiploidy patients based on specific chromosomal gains as detailed in Section 1.3.3.2.1 with many other investigators reporting an association between certain trisomies and prognosis (Enshaei *et al.*, 2021; Moorman *et al.*, 2003; Sutcliffe *et al.*, 2005; Harris *et al.*, 1992; Chang *et al.*, 2024). Moorman *et al.* further reported that age and sex were independent predictors of survival within high hyperdiploidy patients (Moorman *et al.*, 2003). Whilst within *ETV6*::*RUNX1*, it was found that NCI risk affects prognosis (Enshaei *et al.*, 2013).

As well as having good risk features, both subgroups have been shown to respond well to standard chemotherapy regimens, with evidence of sensitivities to common chemotherapeutic agents (Moorman *et al.*, 2003; Sun, Chang and Zhu, 2017; Schrappe *et al.*, 2017; Maloney *et al.*, 2019). For example, Xin Huang *et al.* showed that B-cell ALL patients exhibited a sensitivity to asparaginase (Huang *et al.*, 2024). More specifically, *ETV6*::*RUNX1* were shown to have a relative sensitivity to L-asparaginase when compared to non-*ETV6*::*RUNX1* patients ($p = 0.012$) in a study by Woerden *et al.*, whilst Frost *et al.* showed that these patients were also significantly more sensitive to doxorubicin ($p = 0.001$) and etoposide ($p = 0.001$) (Woerden *et al.*, 2000; Frost *et al.*, 2004). Furthermore, it has been posited that dexamethasone is particularly effective in the treatment of *ETV6*::*RUNX1* patients demonstrated by higher EFS rates at 5-years when compared to prednisolone, as well as excellent outcomes resulting from dexamethasone pulses during maintenance therapy (Bhojwani *et al.*, 2012; Piette *et al.*, 2018). Similarly, it was shown that high hyperdiploidy patients have the ability to accumulate high levels of methotrexate polyglutamates increasing

their sensitivity to the drug (Whitehead *et al.*, 1998; Synold *et al.*, 1994). Kaspers *et al.* found that high hyperdiploidy patients were also sensitive to mercaptopurine ($p < 0.001$), thioguanine ($p = 0.023$), cytarabine ($p = 0.016$), and L-asparaginase ($p = 0.022$) (Kaspers *et al.*, 1995). It was hypothesised that both of these subgroups are sensitive to L-asparaginase due to their low asparagine synthetase expression (Iwamoto *et al.*, 2007). Similarly, Yoshimura *et al.* showed that both subgroups were sensitive to L-asparaginase and prednisolone (Yoshimura *et al.*, 2024).

There is debate on whether *ETV6*::*RUNX1* and high hyperdiploidy are independent predictors of outcome (Ampatzidou *et al.*, 2018). High hyperdiploidy has been shown to be independently prognostic by Hann *et al.* however other studies postulate that it is a heterogeneous subgroup with differing prognosis exemplified by the number of high hyperdiploidy relapses (Hann *et al.*, 2001; Paulsson and Johansson, 2009; Lee *et al.*, 2023a). Similarly, whilst the Dana Farber Cancer Institute Consortium found that *ETV6*::*RUNX1* was not an independent prognostic factor, and other studies have speculated the favourable prognosis of the group due to very late relapses, the Children's Oncology Group and the UK found that the presence of *ETV6*::*RUNX1* was an independent predictor of favourable outcome (Sun, Chang and Zhu, 2017; van Delft *et al.*, 2011; Loh *et al.*, 2006; Rubnitz *et al.*, 2008; Enshaei *et al.*, 2013).

Due to the questions surrounding the true prognostic impact of these good risk genetic groups, they are often not independently used in risk stratification to assign treatment in Europe despite strong evidence of a favourable prognosis within these groups. In the United States of America however, groups have begun treating patients with good risk genetics on low intensity protocols; and with the drive towards identifying patients eligible for treatment de-escalation due to unnecessary toxicities, research groups are starting to assess the viability of that on European protocols (Østergaard *et al.*, 2024; Rubnitz *et al.*, 2008). Østergaard *et al.* assessed the feasibility of treatment de-escalation for *ETV6*::*RUNX1* patients by investigating the outcome of these patients treated on different treatment arms across multiple contemporary trials (Østergaard *et al.*, 2024). However, in the aforementioned study, individual patient data wasn't available/ utilised and historic clinical trials weren't considered. Thus, a horizontal approach to this analysis using consecutive UK clinical trials with individual patient data is undertaken in this thesis on the rationale that optimal treatment elements

exist within these trials due to similarly exceptional outcomes in the good risk genetics subgroups.

3.2 Aims

The aims of this chapter are:

- Determine the survival rates of good risk patients across several historic paediatric clinical trials.

- Investigate the impact of treatment regimen and delayed intensifications on patient outcomes.

- Identify optimal treatment elements for good risk genetic patients which minimise toxicity whilst maintaining satisfactory survival rates.

3.3 Methods

A total of 6716 patients were eligible for analysis in this chapter as outlined in Section 2.3.2. Due to changes to the protocol in 1999, UKALL97 was divided into two groups for this analysis, UKALL97 and UKALL97/99, which were considered and analysed separately. The statistical methods used within this chapter are detailed in Section 2.4. Due to differences in treatment across the trials leading to a general improvement of outcomes over time, the proportional hazards assumption of the Cox proportional hazards model was often violated when stratification by treatment wasn't performed. In these instances, the log-rank p-value has been provided and this information is clearly stated within the text and accompanying tables and figures.

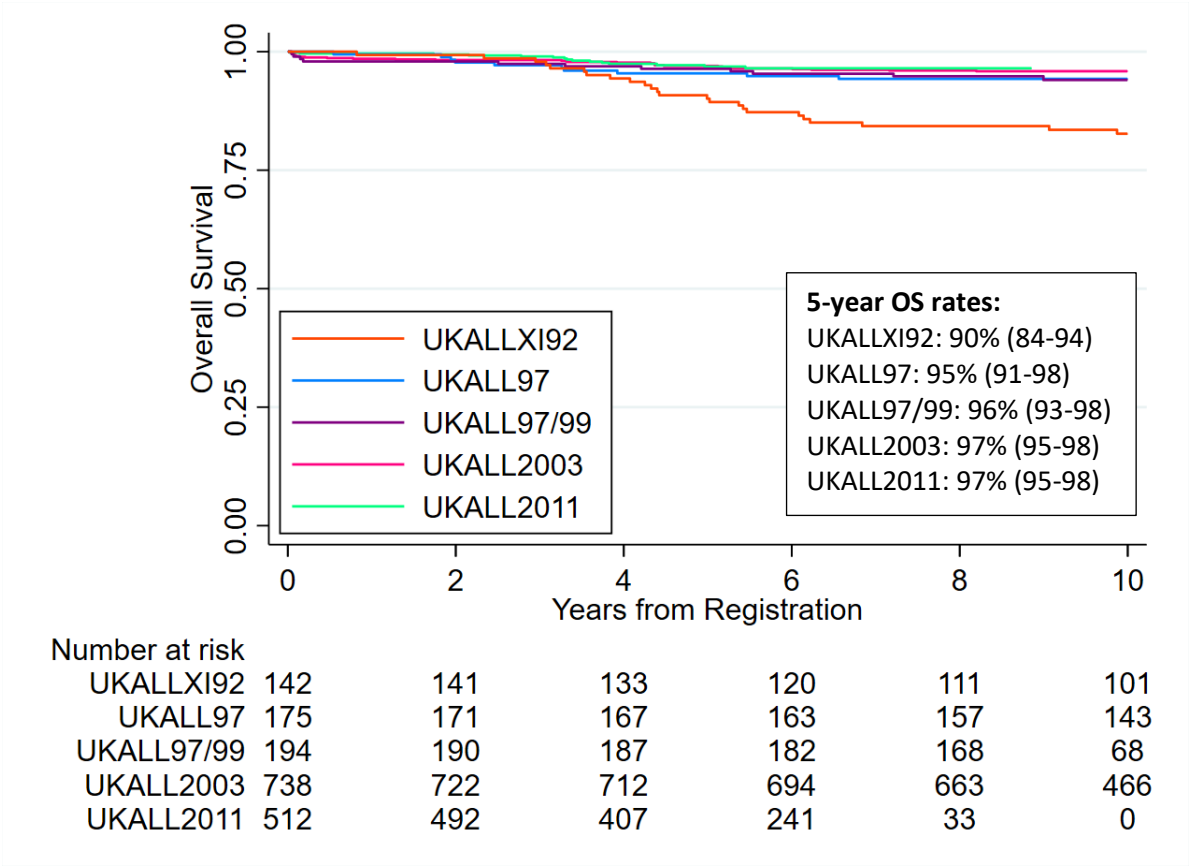
3.4 Results

There is evidence in the literature that outcomes in the B-other subgroup improve over time across trials suggesting improvements in the treatment of these patients. In order to see if this was the case for patients with good risk genetics, outcome across the four most recent paediatric UKALL clinical trials was assessed with the initial focus being at the trial level.

3.4.1 Outcome by trial

3.4.1.1 ETV6::RUNX1

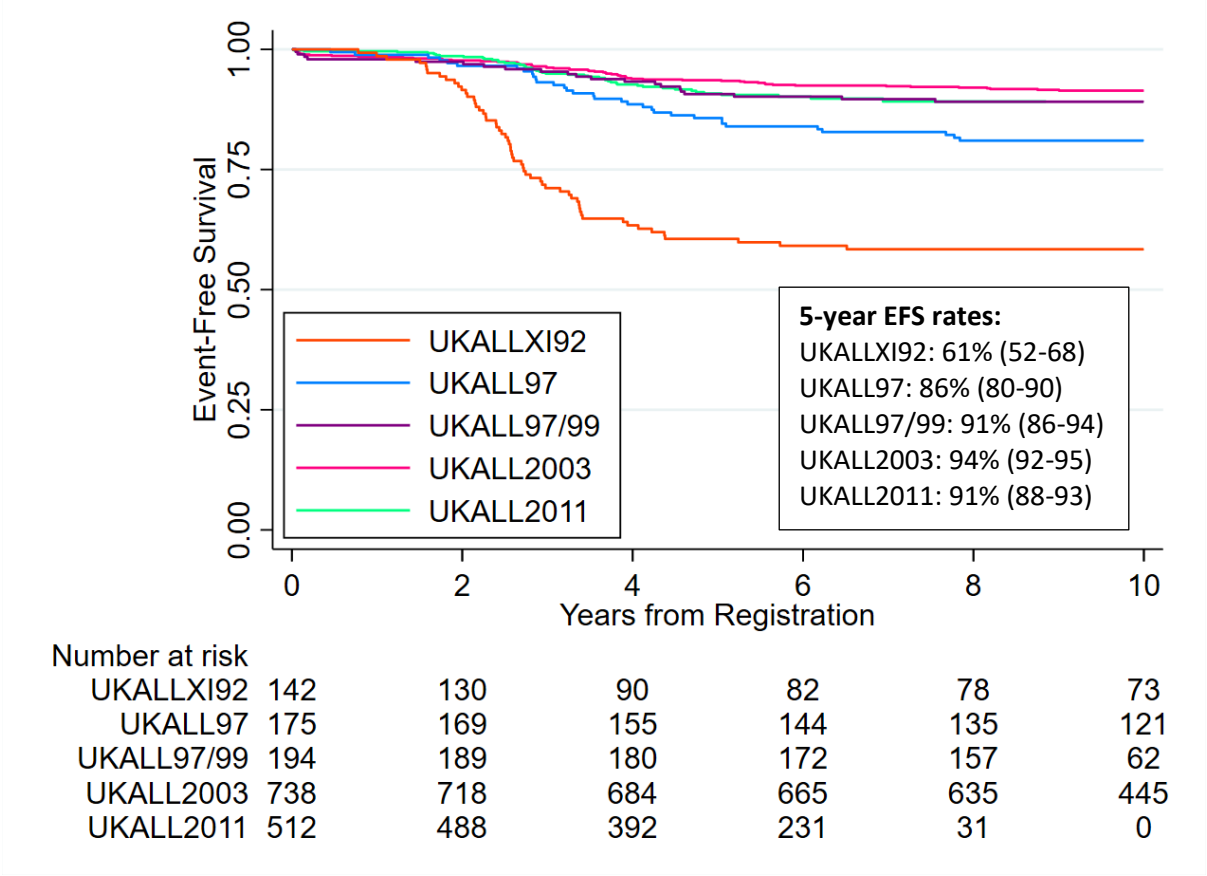
Overall survival of *ETV6*::*RUNX1* patients across the trials remained stable with the exception of UKALLXI92 which had significantly worse survival with a hazard ratio of 3.73, a 95% confidence interval (CI) of (2.34-5.94), and $p < 0.001$ when compared against the other trials combined [Figure 23]. Cure rates for patients treated on the remaining trials was excellent with 5-year survival rates of ~96%. These findings demonstrate that outcome of *ETV6*::*RUNX1* patients has remained stable since UKALL97 despite changes to treatment.



	ALLXI92	ALL97	ALL97/99	ALL2003	ALL2011
Hazard ratio (95% CI), p-value					
OS	2.37 (1.87-3.00), <0.001	1.77 (1.33-2.36), <0.001	1.31 (0.95-1.80), 0.101	1.04 (0.82-1.32), 0.748	1

Figure 23. Kaplan Meier and hazard ratios comparing the overall survival of *ETV6*::*RUNX1* patients on the four most recent paediatric UKALL clinical trials. UKALL2011 is used as the baseline in the Cox proportional hazards models. OS: Overall survival.

ETV6::*RUNX1* patients treated on UKALL97/99, UKALL2003, and UKALL2011 had higher event-free survival compared to the earlier two trials as shown in Figure 24, with 5 year EFS rates >90% for the aforementioned trials and rates of 61% and 86% for UKALLXI92 and UKALL97 respectively. The difference in outcome between the trials stratified by NCI risk and those that weren't was significant, with over twice the hazard for UKALL97 patients and over 5 times the hazard for UKALLXI92 patients with the latter three trials as the baseline (UKALL97: HR = 2.09, 95% CI (1.44-3.04), $p < 0.001$ and UKALLXI92: HR = 5.48, 95% CI (4.02-7.46), $p < 0.001$)). Event-free survival improved with each consecutive trial except in the latter two trials where UKALL2003 had the better outcome with a 5-year EFS rate of 94% (95% CI (92-95)) compared to 91%, 95% CI (88-93) in UKALL2011. Thus, whilst the differences in treatment between the two protocols didn't impair cure rates, it did result in a greater instance of relapses and second tumours. This difference was not significant however with a hazard ratio of 1.37, 95% CI (0.93-2.02), $p = 0.115$.



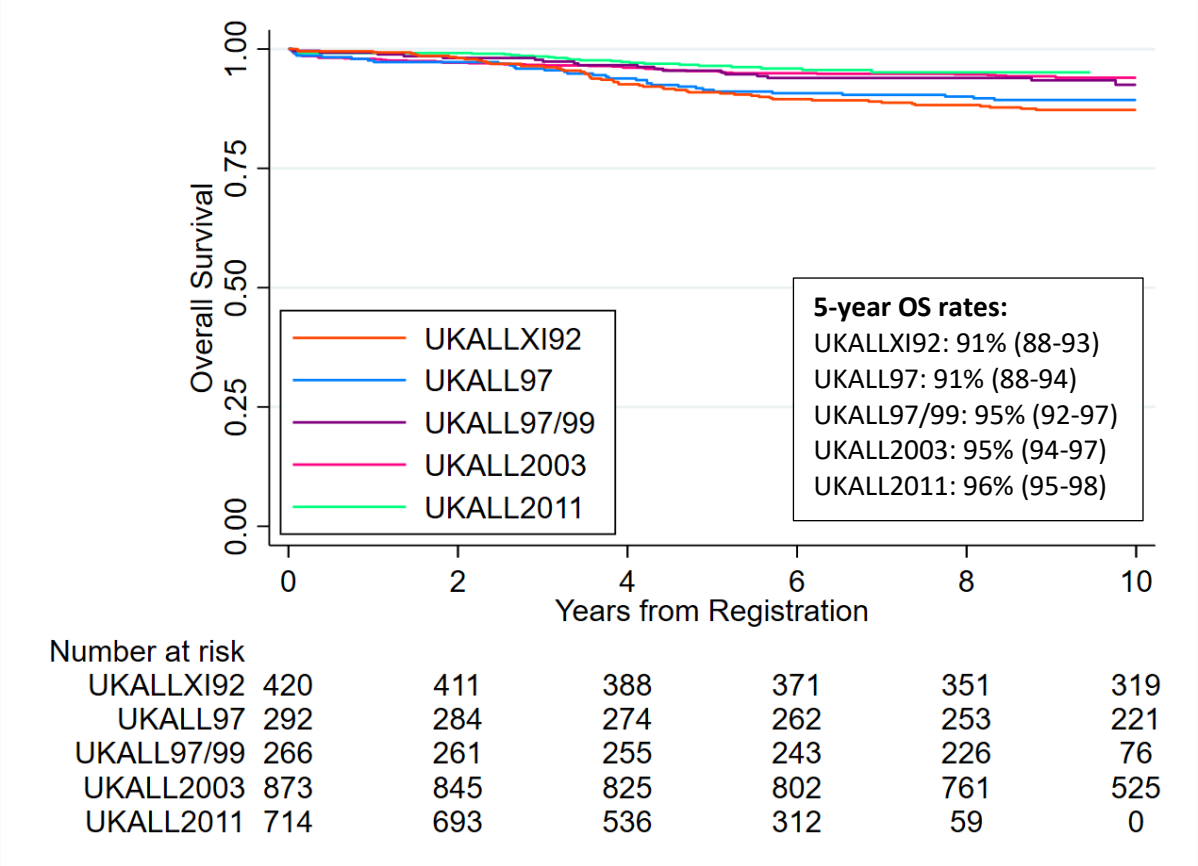
	ALLXI92	ALL97	ALL97/99	ALL2003	ALL2011
Hazard ratio (95% CI), p-value or log-rank p-values					
EFS	Log-rank p<0.001	1.55 (1.27-1.89), <0.001	1.10 (0.88-1.38), 0.396	0.77 (0.65-0.90), 0.002	1

Figure 24. Kaplan Meier and hazard ratios comparing the event-free survival of *ETV6*::*RUNX1* patients on the four most recent paediatric UKALL clinical trials. UKALL2011 is used as the baseline in the Cox proportional hazards models. The log-rank p-value is given in instances where the proportional hazards assumption was violated. EFS: Event-free survival.

3.4.1.2 High hyperdiploidy

Within high hyperdiploidy, overall survival did generally improve with each trial. However, as seen in Figure 25, there are two distinct groups after approximately 4 years with UKALLXI92 and UKALL97 having comparable overall survival of 91% at 5 years whilst the latter three trials also had comparable OS rates of ~95% at 5 years. This difference in outcome seen between the two groups was statistically significant with a hazard ratio of 2.11, 95% CI (1.58-2.82), p <

0.001. This split in the trials coincides with the introduction of NCI risk to assign treatment for patients which suggests that age and white cell count are prognostic within the high hyperdiploidy subgroup.

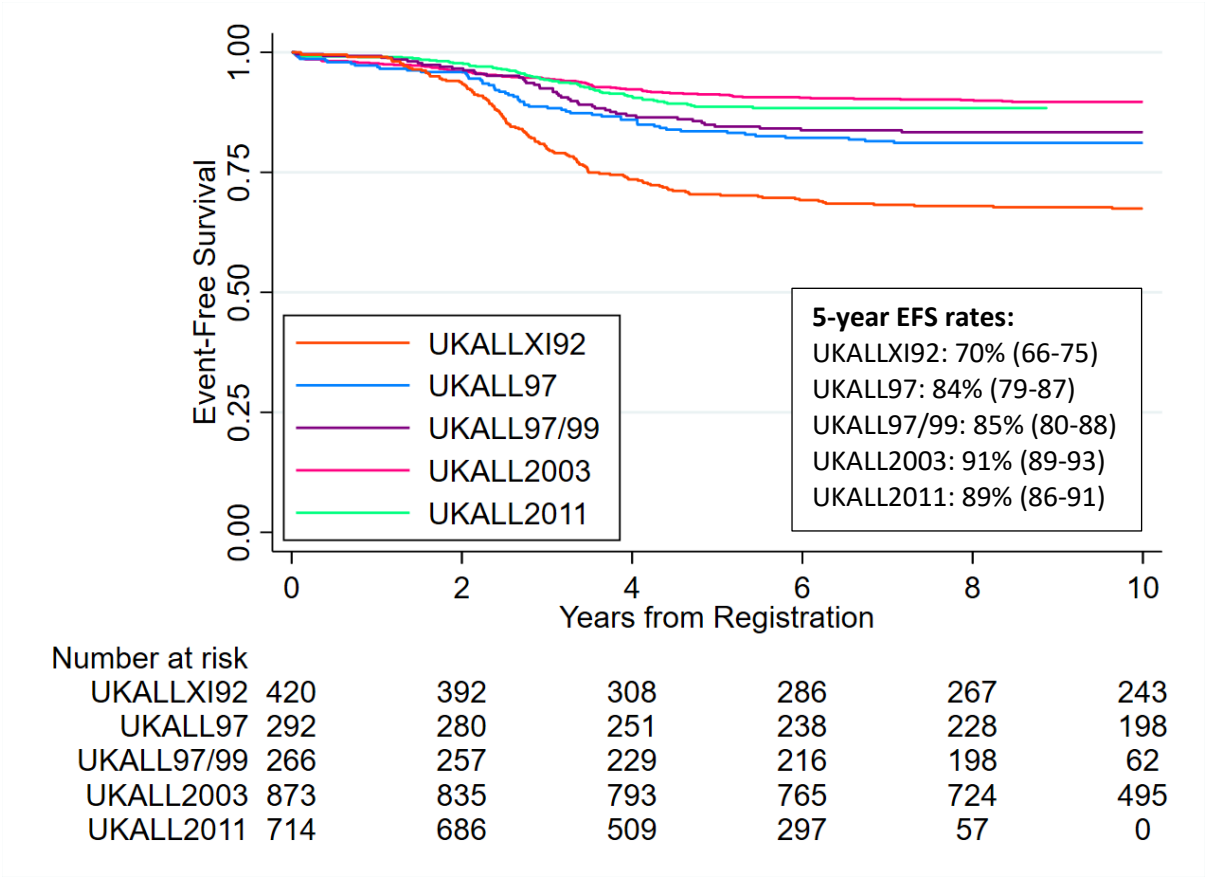


	ALLXI92	ALL97	ALL97/99	ALL2003	ALL2011
Hazard Ratio (95% CI), p					
OS	2.68 (1.68-4.28), <0.001	2.14 (1.27-3.61), 0.004	1.46 (0.80-2.64), 0.215	1.17 (0.73-1.88), 0.523	1

Figure 25. Kaplan Meier and hazard ratios comparing the overall survival of high hyperdiploidy patients on the four most recent paediatric UKALL clinical trials. UKALL2011 is used as the baseline in the Cox proportional hazards models. OS: Overall survival.

In terms of event-free survival, high hyperdiploidy patients treated on UKALLXI92 had the worst outcome of 70% at 5 years with a hazard ratio of 2.81 (95% CI (2.29-3.46), $p < 0.001$) when compared to high hyperdiploidy patients treated on all the other trials combined. It is clear from Figure 26 that there are three distinct groups of survival with UKALL97 and

UKALL97/99 having similar EFS rates (~85% at 5 years) whilst UKALL2003 and UKALL2011 form the group with the best rates (~90% 5-year EFS). These groups have significantly different survival to one another with hazard ratios of 1.98, 95% CI (1.70-2.31), $p < 0.001$ and 0.65, 95% CI (0.56-0.75), $p < 0.001$ for the bottom and top groups respectively when compared to the middle group. This suggests that the measures taken to de-escalate overall treatment across the trials, such as reducing the number of DIs from 3 to 1 and removing HDM and cranial irradiation from CNS-directed therapy, was beneficial in the high hyperdiploidy subgroup. However, UKALL2003 had better event-free survival rates than UKALL2011, thus one could argue that the treatment on UKALL2011 wasn't intensive enough. This difference wasn't statistically significant, however, with a Cox proportional hazards p-value of 0.286. Furthermore, this phenomenon isn't seen in overall survival demonstrating that these additional events aren't affecting the overall cure rate of high hyperdiploidy patients.



	ALLXI92	ALL97	ALL97/99	ALL2003	ALL2011
Hazard Ratio (95% CI), p					
EFS	3.06 (2.30-4.07), <0.001	1.68 (1.19-2.38), 0.003	1.42 (0.98-2.07), 0.066	0.86 (0.63-1.17), 0.334	1

Figure 26. Kaplan Meier and hazard ratios comparing the event-free survival of high hyperdiploidy patients on the four most recent paediatric UKALL clinical trials. UKALL2011 is used as the baseline in the Cox proportional hazards models. EFS: Event-free survival.

3.4.1.3 Representative cohort analysis

In order to ensure the populations were comparable across trials, demographic features were considered for both *ETV6*::*RUNX1* and high hyperdiploidy populations. Table 11 shows the distribution of cases across trials for *ETV6*::*RUNX1* patients. There was no difference in the distribution of *ETV6*::*RUNX1* patients by age or white cell count across trials. There was also no difference in the proportion of males and females on each trial within the *ETV6*::*RUNX1* subgroup with the exception of UKALL97/99 which had a higher proportion of males which was statistically significant with a Pearson χ^2 p-value of 0.002. There was an equal proportion

of patients on regimen B across the trials for which regimens were introduced (UKALL97/99 = 20%, UKALL2003 = 19%, and UKALL2011 = 19%). However, progressively fewer patients were treated on regimen A across the trials (76% vs 70% and 62%) whilst more patients were classified as high risk and treated on regimen C as the trials progressed (5% vs 11% vs 20%). As expected, there was a difference in the distribution of patients by delayed intensification across the trials due to the differences in the protocols.

	Total	ALLXI92	ALL97	ALL97/99	ALL2003	ALL2011	p-value
Total	1761 (100)	142 (8)	175 (10)	194 (11)	738 (42)	512 (29)	
Median Follow-up (years)	9.63	12.99	11.95	9.55	11.03	5.95	
Sex							
Male	963 (55)	75 (53)	92 (53)	126 (65)	400 (54)	270 (53)	
Female	798 (45)	67 (47)	83 (47)	68 (35)	338 (46)	242 (47)	p = 0.047
Age (years)							
Median	4	4.36	4.26	4.04	4.16	4	
1-4	1131 (64)	88 (62)	110 (63)	129 (66)	472 (64)	332 (65)	
5-9	501 (28)	42 (30)	57 (33)	54 (28)	199 (27)	149 (29)	
10-14	111 (6)	12 (8)	8 (5)	10 (5)	55 (7)	26 (5)	
15-19	17 (1)	0 (0)	0 (0)	1 (1)	12 (2)	4 (1)	
≥20	1 (0.06)	0 (0)	0 (0)	0 (0)	0 (0)	1 (0.2)	p = 0.411
White Cell Count (× 10^9/L)							
Median	10.2	14.3	10.7	10.55	10.1	9	
<50	1472 (84)	117 (82)	143 (82)	159 (82)	622 (84)	431 (84)	
≥50	289 (16)	25 (18)	32 (18)	35 (18)	116 (16)	81 (16)	p = 0.851
Regimen							
A	981 (68)	-	-	147 (76)	519 (70)	315 (62)	
B	274 (19)	-	-	38 (20)	140 (19)	96 (19)	
C	189 (13)	-	-	9 (5)	79 (11)	101 (20)	p < 0.001
Delayed Intensifications							
1	779 (45)	0 (0)	0 (0)	0 (0)	267 (36)	512 (100)	
2	766 (44)	65 (49)	36 (21)	194 (100)	471 (64)	0 (0)	
3	205 (12)	68 (51)	137 (79)	0 (0)	0 (0)	0 (0)	p < 0.001

Table 11. Distribution of *ETV6*::*RUNX1* cases across paediatric trials by key demographic, clinical and treatment features. P-values calculated using Pearson χ^2 statistic.

Table 12 shows the distribution of demographic features across the trials for high hyperdiploidy patients. There was no difference in the distribution of sex or white cell count across the trials in the high hyperdiploidy population with ~90% of cases having low white cell counts and the expected ratio of males to females of approximately 1.2:1. There is a significant difference in the distribution of age across the trials, however this is due to the introduction of the enrolment of teenagers and young adults (TYAs) >15 years old onto the more recent trials as there was no significant difference in the distribution when patients aged ≥15 were excluded from the test ($p = 0.284$). Similarly to *ETV6*::*RUNX1*, there were progressively fewer patients treated on the good risk regimen A and more patients treated on regimen C with each consecutive trial.

	Total	ALLXI92	ALL97	ALL97/99	ALL2003	ALL2011	P-value
Total	2565 (100)	420 (16)	292 (11)	266 (10)	873 (34)	714 (28)	
Median Follow-up (years)	9.43	12.65	11.64	9.43	10.93	5.31	
Sex							
Male	1418 (55)	244 (58)	150 (51)	157 (59)	461 (53)	406 (57)	
Female	1147 (45)	176 (42)	142 (49)	109 (41)	412 (47)	308 (43)	P = 0.110
Age (years)							
Median	3.99	4.09	4.05	3.91	3.99	3	
1-4	1617 (63)	260 (62)	186 (64)	168 (63)	538 (62)	465 (65)	
5-9	624 (24)	121 (29)	72 (25)	70 (26)	202 (23)	159 (22)	
10-14	222 (9)	39 (9)	27 (9)	23 (9)	85 (10)	48 (7)	
15-19	92 (4)	0 (0)	7 (2)	5 (2)	42 (5)	28 (5)	
≥20	10 (0.39)	0 (0)	0 (0)	0 (0)	6 (1)	4 (1)	p < 0.001
White Cell Count (× 10^9/L)							
Median	7.2	7.95	7.75	7.05	7.3	6.9	
<50	2331 (91)	382 (91)	256 (88)	244 (92)	793 (91)	656 (92)	
≥50	234 (9)	38 (9)	36 (12)	22 (8)	80 (9)	58 (8)	p = 0.317
Regimen							
A	1166 (63)	-	-	214 (80)	562 (64)	390 (55)	
B	260 (14)	-	-	25 (9)	132 (15)	103 (14)	
C	427 (23)	-	-	27 (10)	179 (21)	221 (31)	p < 0.001
Delayed Intensifications							
1	944 (37)	0 (0)	0 (0)	0 (0)	230 (26)	714 (100)	
2	1174 (47)	204 (53)	61 (21)	266 (100)	643 (74)	0 (0)	
3	406 (16)	183 (47)	223 (79)	0 (0)	0 (0)	0 (0)	p < 0.001

Table 12. Distribution of high hyperdiploidy cases across paediatric trials by key demographic, clinical and treatment features. P-values calculated using Pearson χ^2 statistic.

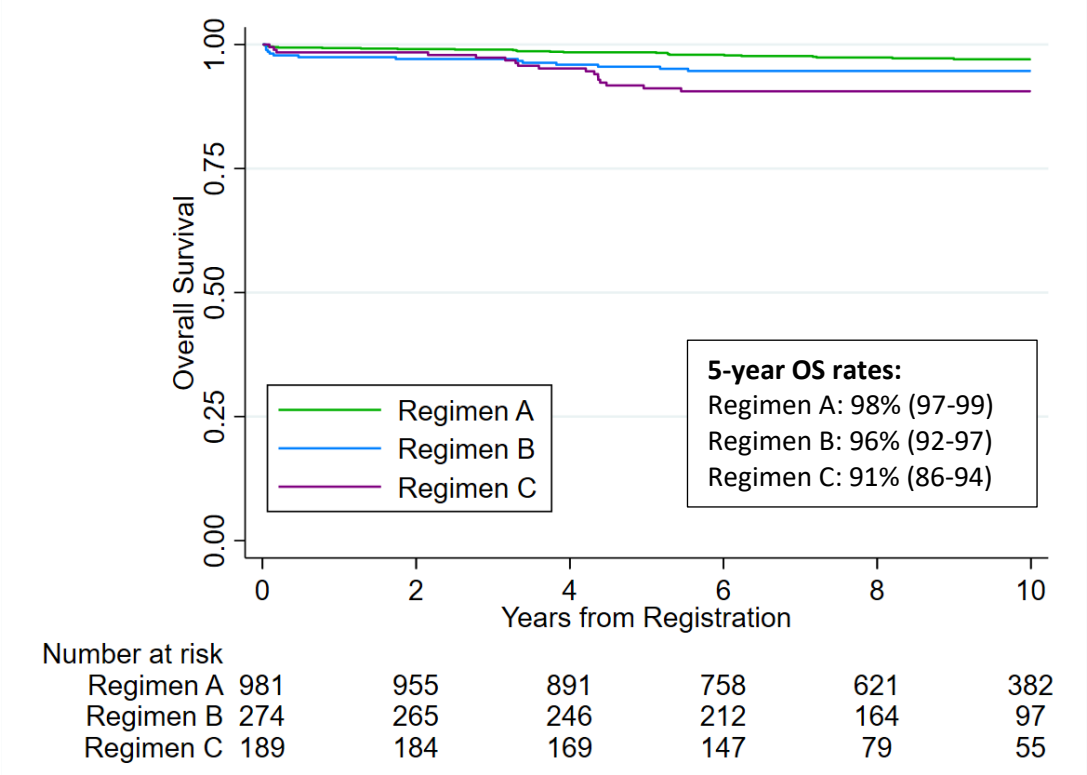
It is clear from the data that more patients from both subgroups were classified and treated as high risk over time as evidenced by the higher proportion of patients on regimen C in the latter trials. As such, the outcome of *ETV6*::*RUNX1* and high hyperdiploidy patients was assessed by regimen to determine if this change was beneficial.

3.4.2 Outcome by regimen

Patients on UKALL97/99, UKALL2003, and UKALL2011 were stratified to regimens A, B, or C based on their NCI risk and genetic abnormality as outlined in Section 2.2. Thus, the analysis in the section was only performed on patients on these trials.

3.4.2.1 ETV6::RUNX1

ETV6::*RUNX1* patients treated on regimen A had the best cure rates with a 5-year overall survival rate of 98% (95% CI (97-99)) whilst regimen C patients had the lowest cure rates at 91% (95% CI (86-94)) at the same time point [Figure 27]. The difference in overall survival was significant between regimen A compared to both regimen B (log-rank: p = 0.0406) and regimen C (Cox: p < 0.001), however there was no significant difference in survival between regimen B and Regimen C (Cox: p = 0.111).



	Regimen A	Regimen B	Regimen C
Hazard Ratio (95% CI), p			
Overall Survival	1	1.95 (1.02-3.73), 0.045	3.46 (1.88-6.38), <0.001

Figure 27. Kaplan Meier and hazard ratios comparing the difference in overall survival of *ETV6*::*RUNX1* patients stratified by regimen. Regimen A is used as the baseline in the Cox proportional hazards models. OS: Overall survival.

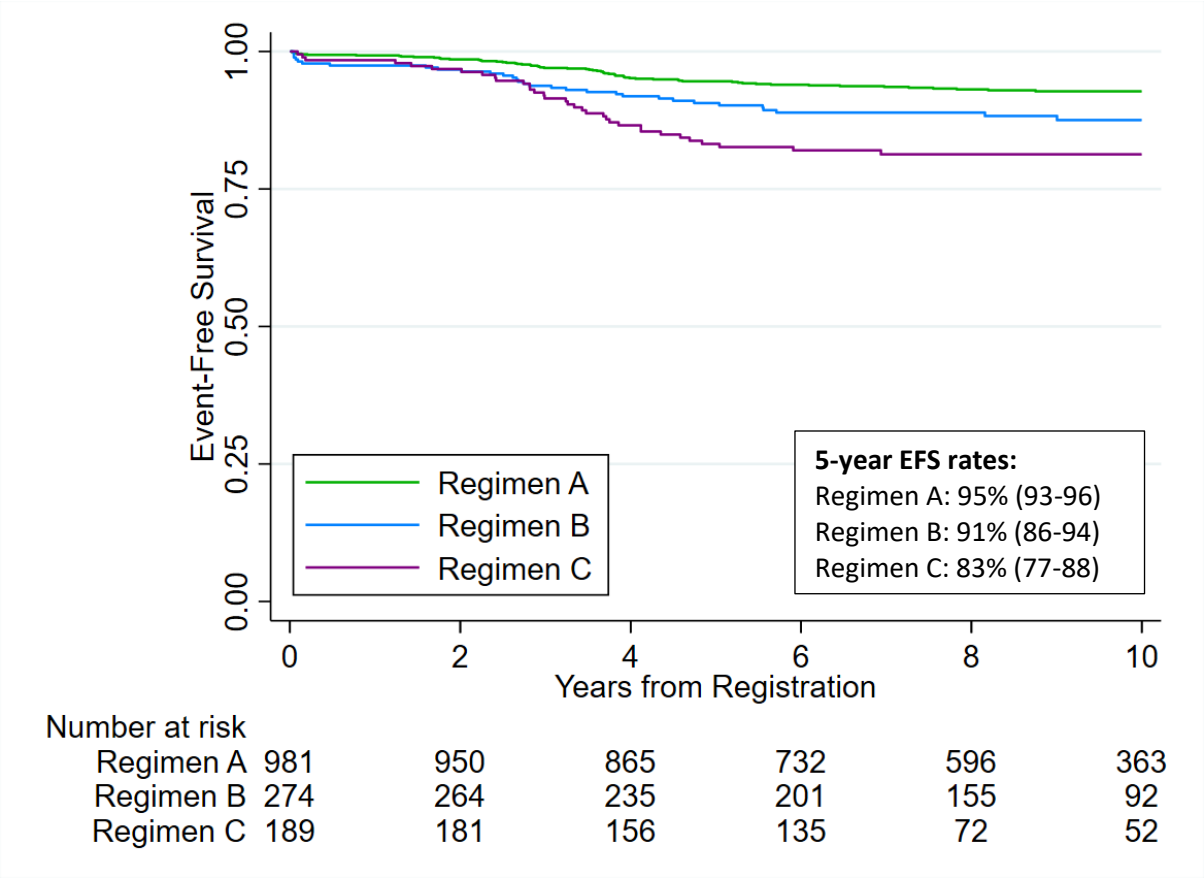
It is clear from Table 13 that cure rates were similar across the three trials for *ETV6*::*RUNX1* patients treated on regimens A and B at ~99% and ~95% respectively. As such, there was no significant difference in the hazards of patients treated on UKALL97/99 or UKALL2003 when compared to UKALL2011 with p-values of 0.075 and 0.095 respectively for regimen A and 0.654 and 0.618 for regimen B. For patients assigned to regimen C however, survival rates were far inferior for patients treated on UKALL97/99 with a 5-year EFS rate of 67% compared to 92% and 93% for patients treated on UKALL2003 and UKALL2011 respectively. There was no significant difference in the hazards for patients who received regimen C on UKALL2003 and UKALL2011 (HR: 0.87, 95% CI (0.30-2.52), p = 0.804), however there was a significantly greater hazard for regimen C patients on UKALL97/99 compared to UKALL2011 (HR: 4.89, 95% CI (1.30-18.44), p = 0.019.

	UKALL97/99	**UKALL2003**	**UKALL2011**
5-year overall survival rates (95% CI)			
Regimen A	99% (95-99.7)	98% (96-99)	99% (97-99.8)
Regimen B	95% (81-99)	96% (92-98)	95% (87-98)
Regimen C	67% (28-88)	92% (84-97)	93% (85-96)
Overall survival hazard ratio (95% CI), p-value			
Regimen A	4.33 (0.86-21.73), p = 0.075	3.52 (0.80-15.44), p =0.095	1
Regimen B	1.39 (0.33-5.83), p = 0.654	0.74 (0.22-2.43), p = 0.618	1
Regimen C	4.89 (1.30-18.44), p = 0.019	0.87 (0.30-2.52), p = 0.804	1

Table 13. 5-year overall survival rates and hazard ratios for *ETV6*::*RUNX1* patients stratified by regimen across trials. UKALL2011 is used as the baseline in the cox proportional hazards models.

For event-free survival, *ETV6*::*RUNX1* patients again saw an improvement in outcomes sequentially from regimen C through to regimen A as seen in Figure 28. This difference was more pronounced than in overall survival, with 17% of patients having an event by 5 years for

regimen C compared to only 5% and 9% for regimens A and B respectively. The difference in event-free survival was significant between all three regimens, with regimen A having approximately half the risk of and event compared to regimen B patients (HR: 0.57, 95% CI (0.37-0.88), p = 0.01), whilst regimen C patients had a ~64% increased risk (HR: 1.64, 95% CI (1.01-2.67), p = 0.047) when compared to regimen B.



	Regimen A	Regimen B	Regimen C
Hazard Ratio (95% CI), p			
Event-Free Survival	0.57 (0.37-0.88), 0.01	1	2.87 (1.89-4.34), <0.001

Figure 28. Kaplan Meier and hazard ratios comparing the event-free survival of *ETV6*::*RUNX1* patients stratified by regimen. Regimen A is used as the baseline in the Cox proportional hazards models. EFS: Event-free survival.

There was no significant difference in EFS for *ETV6*::*RUNX1* patients on any regimen across the trials. Patients treated on regimen A had identical 5-year EFS rates of 95% in each trial and no significant difference in hazard with p-values of 0.656 and 0.929 for UKALL97/99 and UKALL2003 compared to UKALL2011 [Table 14]. For regimen B, patients treated on

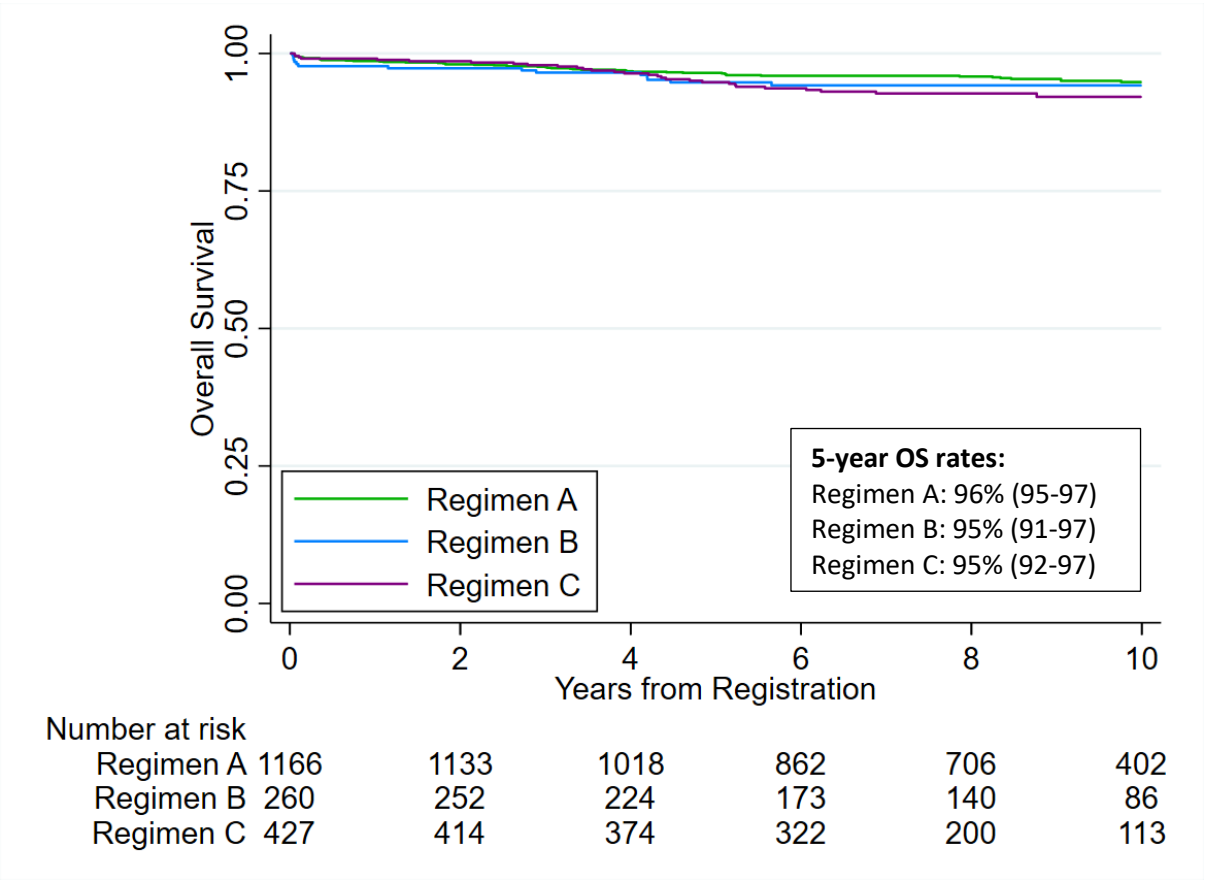
UKALL2003 had the most superior 5-year EFS rates at 94% compared to 84% and 88% for UKALL97/99 and UKALL2011 respectively. This improvement was not significant however, with a hazard ratio of 0.68, 95% CI (0.30-1.53), p = 0.350 compared to UKALL2011. Similarly, the hazard for UKALL97/99 was not significant different to UKALL2011 at 1.17 (p = 0.754). Although patients who received regimen C treatment on UKALL97/99 had a much inferior survival to patients on UKALL2003 and UKALL2011, the difference in hazard between UKALL2011 and UKALL97/99 was non-significant (HR: 2.95, 95% CI (1.00-8.67), p = 0.05) as was the hazard between UKALL2011 and UKALL2003 (p = 0.298).

	UKALL97/99	**UKALL2003**	**UKALL2011**
5-year event-free survival rates (95% CI)			
Regimen A	95% (89-97)	95% (92-96)	95% (91-97)
Regimen B	84% (68-93)	94% (88-97)	88% (79-94)
Regimen C	56% (20-80)	87% (78-93)	82% (73-89)
Event-free survival hazard ratio (95% CI), p-value			
Regimen A	1.19 (0.56-2.53), p = 0.656	1.03 (0.57-1.87), p =0.929	1
Regimen B	1.17 (0.43-3.20), p = 0.754	0.68 (0.30-1.53), p = 0.350	1
Regimen C	2.95 (1.00-8.67), p = 0.05	0.67 (0.32-1.42), p = 0.298	1

Table 14. 5-year event-free survival rates and hazard ratios for *ETV6*::*RUNX1* patients stratified by regimen across trials. UKALL2011 is used as the baseline in the Cox proportional hazards models.

3.4.2.2 High hyperdiploidy

Unlike the *ETV6*::*RUNX1* subgroup, there was no difference in overall survival in high hyperdiploidy patients between any of the regimens with almost identical 5-year OS rates (96% vs 95% vs 95%) as evidenced in Figure 29. Hazard ratios confirm the lack of significant difference between both regimen B (HR = 1.19, 95% CI (0.66-2.15), p = 0.555) and regimen C (HR = 1.50, 95% CI (0.96-2.36), p = 0.077) compared to regimen A.



	Regimen A	Regimen B	Regimen C
Hazard Ratio (95% CI), p			
Overall Survival	1	1.19 (0.66-2.15), 0.555	1.50 (0.96-2.36), 0.077

Figure 29. Kaplan Meier and hazard ratios comparing the overall survival of high hyperdiploidy patients stratified by regimen. Regimen A is used as the baseline in the Cox proportional hazards models. OS: Overall survival.

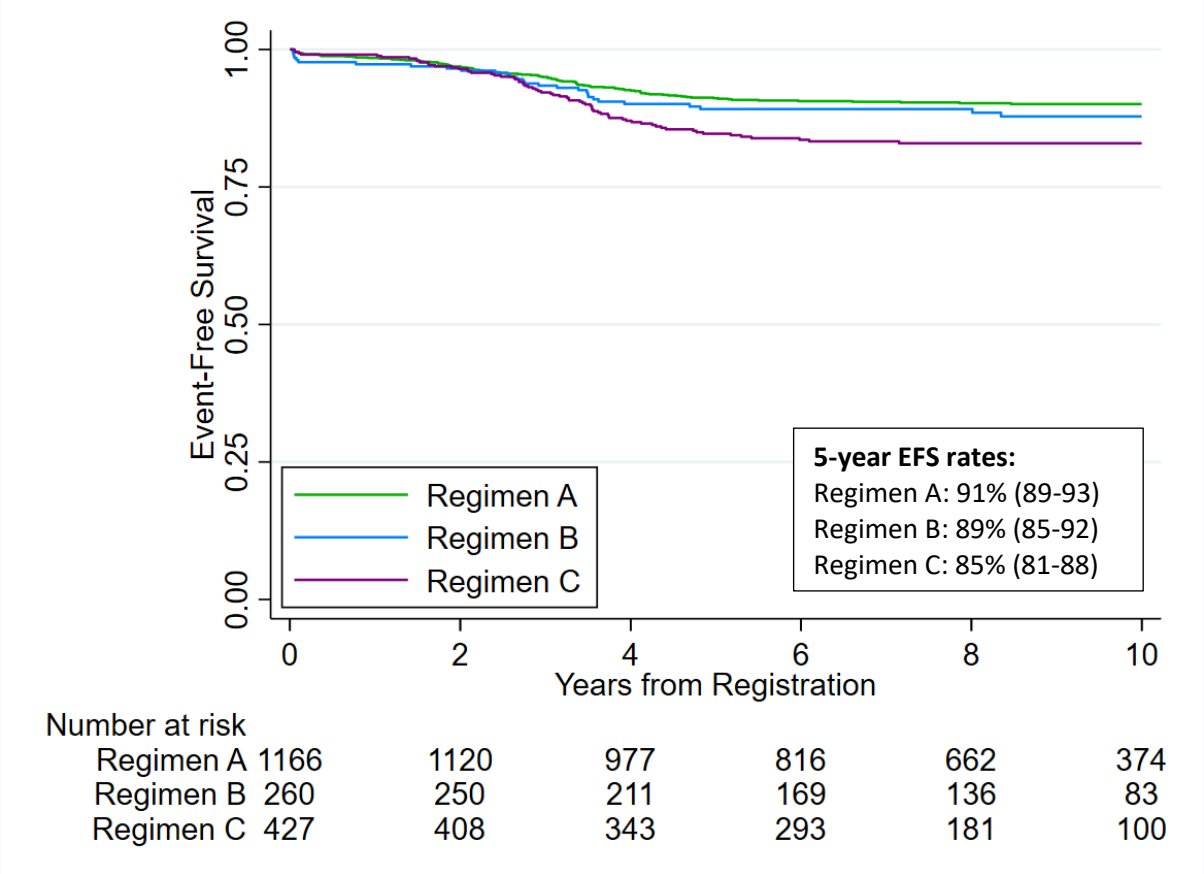
For high hyperdiploidy, cure rates were comparable by regimen across the trials. Patients stratified to regimen A had 5-year OS rates of 96%, 95%, and 94% on UKALL97/99, UKALL2003, and UKALL2011 respectively [Table 15]. Within the regimen B and C subgroups, patients treated on UKALL2011 had the worst outcomes overall, with OS rates of 87% and 83% at 5 years compared to rates >90% for patients on UKALL97/99 and UKALL2003. This difference was not significant though as is evidenced by the hazard ratios for regimen B (HR: 2.32 (0.39-13.93), p = 0.359 and HR: 2.03 (0.54-7.54), p = 0.292. In regimen C the hazard ratio for

UKALL97/99 was 2.65 (0.83-8.51), p = 0.101 compared to UKALL2011 whilst the difference in outcome for UKALL2003 was also non-significant by log-rank test (p = 0.247) which was performed as the proportional hazard assumption was violated.

	UKALL97/99	UKALL2003	UKALL2011
5-year overall survival rates (95% CI)			
Regimen A	96% (92-98)	97% (95-98)	94% (91-96)
Regimen B	96% (75-99)	93% (87-96)	87% (82-91)
Regimen C	93% (74-98)	93% (88-96)	83% (77-88)
Overall survival hazard ratio (95% CI), p-value			
Regimen A	1.31 (0.60-2.86), p = 0.5	1.01 (0.51-1.99), p =0.982	1
Regimen B	2.32 (0.39-13.93), p = 0.359	2.03 (0.54-7.54), p = 0.292	1
Regimen C	2.65 (0.83-8.51), p = 0.101	Log rank p-value = 0.247	1

Table 15. 5-year overall survival rates and hazard ratios for high hyperdiploidy patients stratified by regimen across trials. UKALL2011 is used as the baseline in the Cox proportional hazards models. The log-rank p-value is given in instances where the proportional hazards assumption was violated.

There was also no significant difference between regimens A and B in event-free survival (HR = 1.22, 95% CI (0.81-1.83), p = 0.350) with 5-year EFS rates of 91% and 89% respectively [Figure 30]. Furthermore, there was no difference in outcomes between regimen B and regimen C with survival rates of 89% and 85% at 5-years (HR = 1.45, 95% CI (0.94-2.24), p = 0.092). There was, however, a statistically significant difference seen between regimen A and regimen C (HR = 1.77, 95% CI (1.31-2.39), p < 0.001). It is clear from the lack of difference in cure rates however that these events were salvageable.



	Regimen A	Regimen B	Regimen C
Hazard Ratio (95% CI), p			
Event-Free Survival	1	1.22 (0.81-1.83), 0.350	1.77 (1.31-2.39), <0.001

Figure 30. Kaplan Meier and hazard ratios comparing the event-free survival of high hyperdiploidy patients stratified by regimen. Regimen A is used as the baseline in the Cox proportional hazards models. EFS: Event-free survival.

Comparing event-free survival of high hyperdiploidy patients by regimen across trials, it is clear from Table 16 that the outcomes were similar on each trial. The 5-year EFS rates range from 85-93% on regimen A, 84-92% on regimen B, and 78%-86% on regimen C. These differences are mostly non-significant, except that patients treated on regimen A of UKALL97/99 had a significantly higher hazard than UKALL2011 patients at 1.85, 95% CI (1.11-3.07), p = 0.017.

	UKALL97/99	UKALL2003	UKALL2011
5-year event-free survival rates (95% CI)			
Regimen A	85% (80-90)	93% (90-95)	92% (88-94)
Regimen B	84% (62-94)	92% (85-95)	87% (77-92)
Regimen C	78% (57-89)	86% (80-90)	84% (79-89)
Event-free survival hazard ratio (95% CI), p-value			
Regimen A	1.85 (1.11-3.07), p = 0.017	0.98 (0.61-1.57), p =0.93	1
Regimen B	1.10 (0.35-3.44), p = 0.875	0.67 (0.30-1.51), p = 0.336	1
Regimen C	1.63 (0.72-3.69), p = 0.238	1.01 (0.61-1.67), p = 0.962	1

Table 16. 5-year event-free survival rates and hazard ratios for high hyperdiploidy patients stratified by regimen across trials. UKALL2011 is used as the baseline in the Cox proportional hazards models.

Overall, these findings suggest that the stratification of more patients to Regimen C in the latter trials was not beneficial for *ETV6*::*RUNX1* or high hyperdiploidy patients, who largely saw no difference in cure rates by regimen across trials; thus implying the additional therapy was unnecessary for those patients as the regimen A and B survival rates did not improve with the additional patients moving to regimen C.

3.4.2.3 Representative cohort analysis

In order to confirm the legitimacy of these findings, demographics of the two subgroups were compared across regimen. Within *ETV6*::*RUNX1*, there was no significant difference in sex across the three regimens [Table 17]. As expected, there was a difference in age and white cell count as these were the criteria for the stratification by regimen. We can see that the median age across the regimens was similar at ~4 years old which coheres with the fact that *ETV6*::*RUNX1* patients are associated with a younger age.

	Total	Regimen A	Regimen B	Regimen C	P-value
Total	1444 (100)	981 (68)	274 (19)	189 (13)	
Median Follow-up (years)	8.92	9.16	8.81	7.8	
Sex					
Male	796 (55)	551 (56)	138 (50)	107 (57)	
Female	648 (45)	430 (44)	136 (50)	82 (43)	p = 0.211
Age (years)					
Median	4	4	4.3	4	
1-4	933 (65)	658 (67)	150 (55)	125 (66)	
5-9	402 (28)	323 (33)	38 (14)	41 (22)	
10-14	91 (6)	0 (0)	70 (26)	21 (11)	
15-19	17 (1)	0 (0)	15 (5)	2 (1)	
≥20	1 (0.07)	0 (0)	1 (0.36)	0 (0)	p < 0.001
White Cell Count ($\times 10^9$/L)					
Median	9.85	7.8	66.5	13.3	
<50	1212 (84)	979 (99.8)	81 (30)	152 (80)	
≥50	232 (16)	2 (0.2)	193 (70)	37 (20)	p < 0.001

Table 17. Distribution of *ETV6*::*RUNX1* cases across regimens by key demographic features.

P-values calculated using Pearson χ^2 statistic.

Similarly to *ETV6*::*RUNX1* patients, there is no difference in sex by regimen within the high hyperdiploidy subgroup, yet there is a significant difference in the proportions of age and white cell count. However unlike within *ETV6*::*RUNX1*, the median age is much higher for regimen B compared to regimen A (11.27 vs 3.38 years) whilst regimen C patients have a similar median age to regimen A at 4 years old. Despite the significant difference in the proportion of low-risk white cell counts between regimens A and B, the median white cell counts in these groups are similar at 6.95 and 6.3. This is shown in Table 18 below.

	Total	Regimen A	Regimen B	Regimen C	p-value
Total	1853 (100)	1166 (63)	260 (14)	427 (23)	
Median Follow-up (years)	8.62	8.97	8.81	8	
Sex					
Male	1024 (55)	645 (55)	144 (55)	235 (55)	
Female	829 (45)	521 (45)	116 (45)	192 (45)	p = 0.994
Age (years)					
Median	3.91	3.38	11.27	4	
1-4	1171 (63)	846 (73)	79 (30)	246 (58)	
5-9	431 (23)	320 (27)	12 (5)	99 (23)	
10-14	156 (8)	0 (0)	102 (39)	54 (13)	
15-19	85 (5)	0 (0)	61 (23)	24 (6)	
≥20	10 (0.54)	0 (0)	6 (2)	4 (1)	p < 0.001
White Cell Count ($\times 10^9$/L)					
Median	7.1	6.95	6.3	8	
<50	1693 (91)	1165 (99.9)	166 (64)	362 (85)	
≥50	160 (9)	1 (0.09)	94 (36)	65 (15)	p < 0.001

Table 18. Distribution of high hyperdiploidy cases across regimens by key demographic features. P-values calculated using Pearson χ^2 statistic.

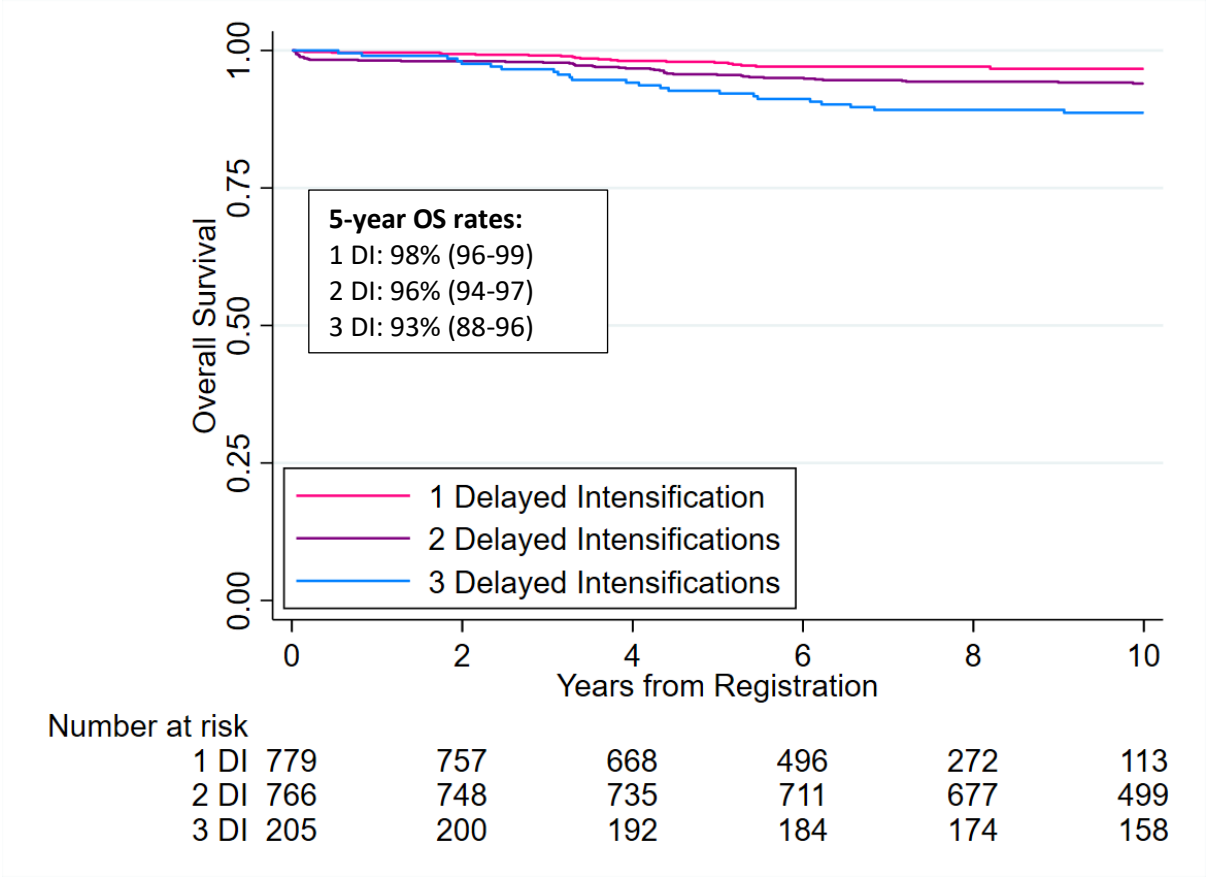
3.4.3 Outcome by delayed intensifications

Patients treated on the clinical trials considered in this thesis received 1, 2, or 3 delayed intensifications. This was either due to randomisations or because everyone on a trial received the same number of intensifications as outlined in Section 2.2.

3.4.3.1 ETV6::RUNX1

As shown in Figure 31, the 5-year overall survival rates for *ETV6*::*RUNX1* patients are similar for the three groups ranging from 93% - 98%. However, there is significant increased risk in the two and three DI groups compared to one delayed intensification (HR = 1.84, 95% CI (1.10-3.10), p = 0.021 and HR = 3.61, 95% CI (2.00-6.50), p < 0.001 respectively). There is also

significantly inferior survival in the third intensification subgroup when compared to two intensifications with a hazard ratio of 1.96 (95% CI (1.19-3.20), p = 0.008).



	1 DI	2 DI	3 DI
Hazard Ratio (95% CI), p			
Overall Survival	1	1.84 (1.10-3.10), 0.021	3.61 (2.00-6.50), <0.001

Figure 31. Kaplan Meier and hazard ratios comparing the overall survival of *ETV6*::*RUNX1* patients stratified by delayed intensification. One delayed intensification is used as the baseline in the Cox proportional hazards models. DI: delayed intensification, OS: overall survival.

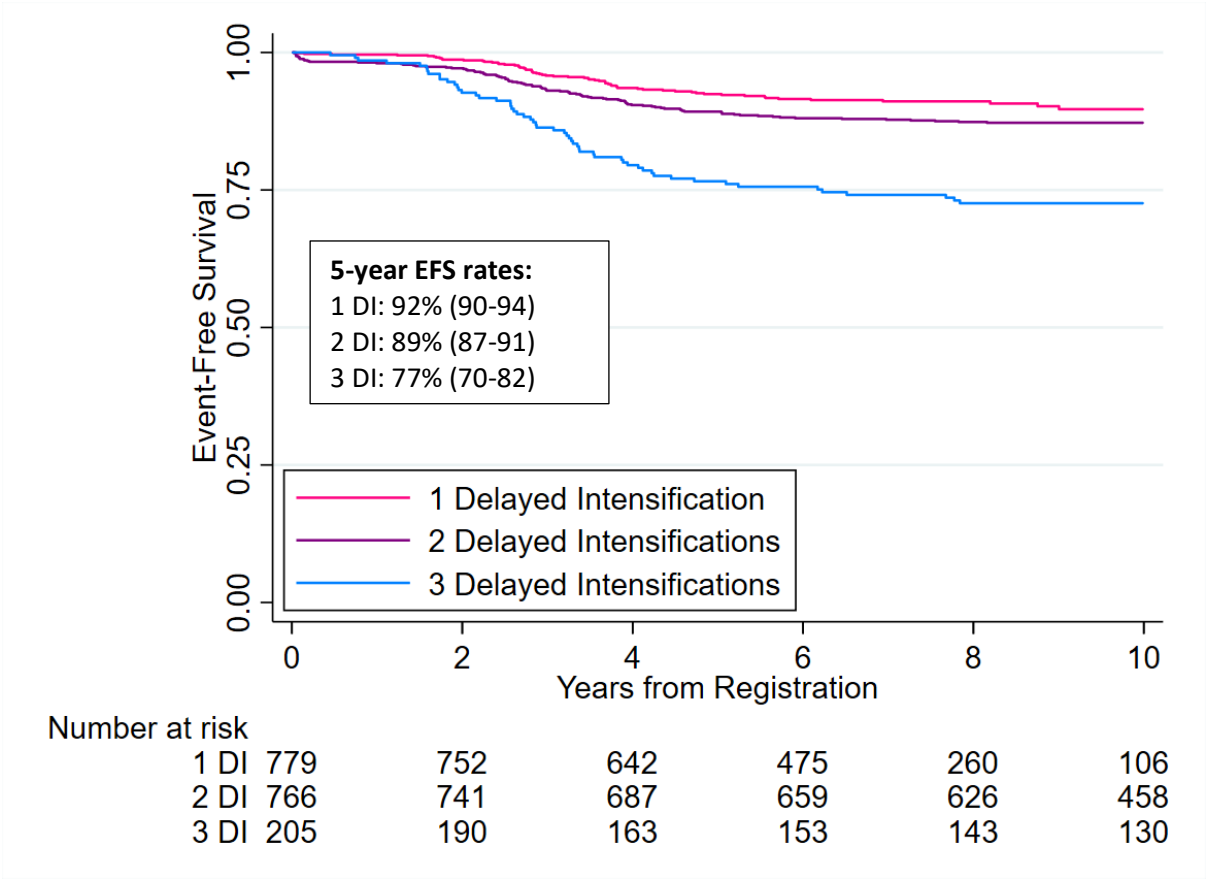
When comparing *ETV6*::*RUNX1* patients stratified by number of DIs received across trial, it is evident from Table 19 that there was no difference in outcome between the patients on UKALL2003 and UKALL2011 who received 1 DI with 5-year OS rates of 99% and 97% respectively. The hazard ratio for 1 DI patients on UKALL2003 compared to UKALL2011 was 0.54, 95% CI (0.20-1.45), p = 0.221 showing that there was no statistically significant

difference in the risk of death in these groups. Within 2 delayed intensifications, patients treated on UKALLXI92 had the most inferior outcome with 5-year OS rates of 89% compared to 100%, 96%, and 96% for UKALL97, UKALL97/99, and UKALL2003 respectively. This difference was statistically significant as this group had a hazard of 3.07, 95% CI (1.47-6.41), p = 0.003 compared to UKALL2003 which was the group with the most similar outcome. Patients treated on UKALLXI92 who received 3 DIs performed significantly worse than those on UKALL97, with a hazard of 2.42, 95% CI (1.08-5.41), p = 0.032. When comparing 5-year survival rates of the two delayed intensification groups on UKALL2003, it is evident that *ETV6*::*RUNX1* patients receiving 1 DI had superior survival to those receiving 2 DI at 99% vs 96%. However this difference was not significant with a hazard ratio of 2.31, 95% CI (0.94-5.65), p = 0.066 for the 2 DI group. Interestingly, patients treated on UKALLXI92 had comparable survival rates (~90%) for patients treated with 2 or 3 DIs, whilst patients treated on UKALL97 had superior 5 year OS rates in the 2 DI group at 100% vs 94%.

	UKALLXI92	**UKALL97**	**UKALL97/99**	**UKALL2003**	**UKALL2011**
5-year overall survival rates (95% CI)					
1 DI	-	-	-	99% (97-99.8)	97% (95-98)
2 DI	89% (79-95)	100%	96% (93-98)	96% (93-97)	-
3 DI	90% (79-95)	94% (89-97)	-	-	-
Overall survival hazard ratio (95% CI), p-value					
1 DI	-	-	-	0.54 (0.20-1.45), p =0.221	1
2 DI	3.07 (1.47-6.41), 0.003	-	1.26 (0.63-2.54), p = 0.509	1	-
3 DI	2.42 (1.08-5.41), 0.032	1	-	-	-

Table 19. 5-year overall survival rates and hazard ratios for *ETV6*::*RUNX1* patients stratified by delayed intensification across trials. The latest trial available is used as the baseline in the Cox proportional hazards models. DI: delayed intensification.

For *ETV6*::*RUNX1* patients in the event-free survival setting, there was once again a difference in outcome between the delayed intensifications with 5-year EFS rates of 92%, 89%, and 77% for 1, 2, and 3 delayed intensifications respectively [Figure 32]. When compared to 1 delayed intensification, both 2 and 3 DIs had significantly inferior survival with p-values of 0.043 and $p < 0.001$ respectively. The log-rank p-value is presented for the comparison between 1 and 2 delayed intensifications due to violating the proportional hazards assumption. The inferior survival seen in the three DI group could be due to the fact that only patients on the earlier trials (who had lower survival rates compared to patients on the more recent trials) received three delayed intensifications; whilst only patients on the two most recent trials were eligible to receive 1 intensification. Investigation by trial was performed to assess this.



	1 DI	2 DI	3 DI
Hazard Ratio (95% CI), p			
Event-Free Survival	1	Log-rank p-value = 0.043	3.28 (2.30-4.67), <0.001

Figure 32. Kaplan Meier and hazard ratios comparing the event-free survival of *ETV6*::*RUNX1* patients stratified by delayed intensification. One delayed intensification is used as the baseline in the Cox proportional hazards models. The log-rank p-value is given in instances where the proportional hazards assumption was violated. DI: delayed intensification, EFS: event-free survival.

ETV6::*RUNX1* patients who received 1 delayed intensification on UKALL2003 had significantly better event-free survival than those on UKALL2011, with a hazard ratio of 0.56 (0.32-0.99), p = 0.047 [Table 20]. Patients who received 2 DIs had comparable outcomes on UKALL97, UKALL97/99, and UKALL2003 with survival rates ~92%, whilst patients on UKALLXI92 had significantly inferior EFS with 5-year rates of 57% and a hazard ratio of 6.32, 95% CI (3.94-10.13), p < 0.001 compared to UKALL2003. Patients who received 3 DIs on UKALLXI92 also

had significantly worse survival when compared to those treated on UKALL97 with a hazard ratio of 2.07, 95% CI (1.23-3.47), p = 0.006. For patients treated on UKALL2003, those who received 1 delayed intensification performed slightly better than patients who received 2 delayed intensifications (95% vs 93% EFS at 5-years) but this difference was not significant with a p-value of 0.35. As in overall survival, patients treated on UKALL97 had improved survival when receiving 2 DIs compared to 3 (94% vs 83%) whilst patients on UKALLXI92 had inferior EFS when receiving 2 DIs (57% vs 63%). Whilst these differences were not significant, with p-values of 0.123 and 0.594 respectively, this was an interesting observation, so further investigation into this difference was carried just in the patients who were randomised to 2 or 3 delayed intensifications on these trials and is described in Section 3.4.4.

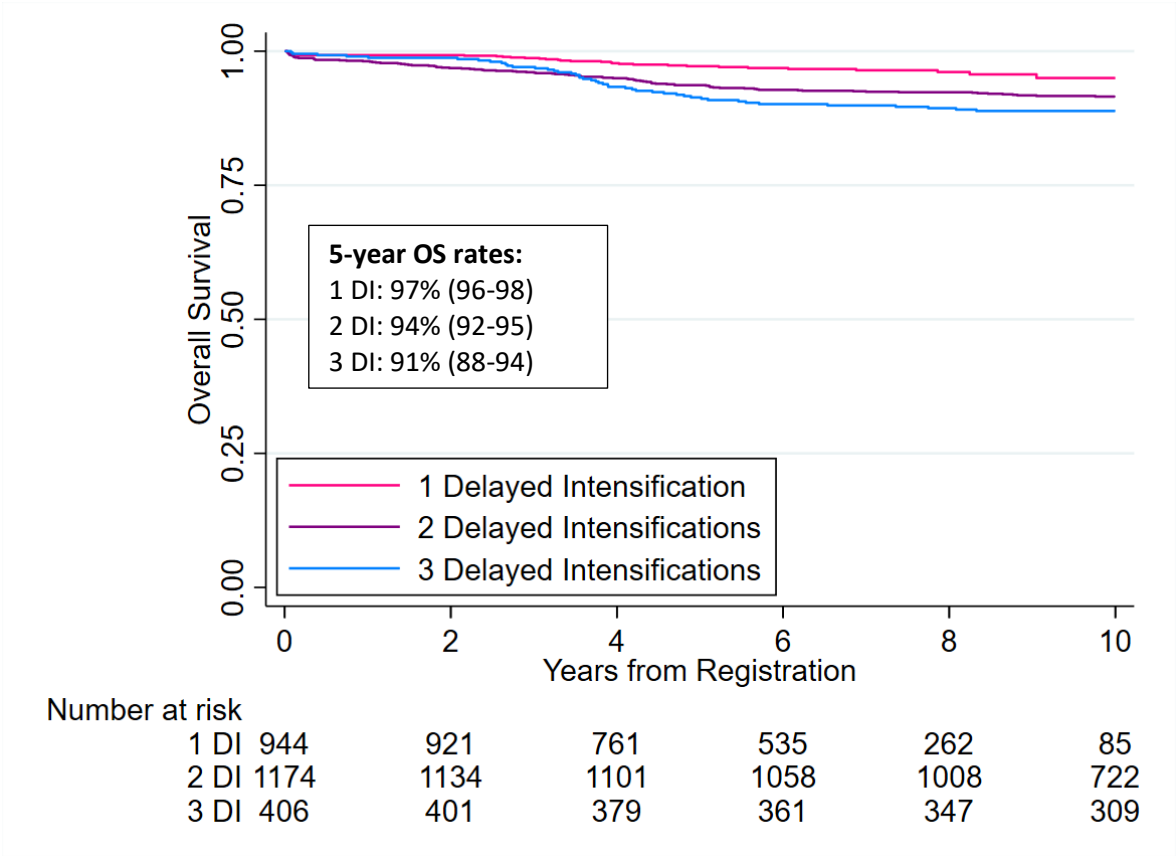
	UKALLXI92	**UKALL97**	**UKALL97/99**	**UKALL2003**	**UKALL2011**
5-year event-free survival rates (95% CI)					
1 DI	-	-	-	95% (92-97)	91% (88-93)
2 DI	57% (44-68)	94% (80-99)	91% (86-94)	93% (90-95)	-
3 DI	63% (51-73)	83% (76-89)	-	-	-
Event-free survival hazard ratio (95% CI), p-value					
1 DI	-	-	-	0.56 (0.32-0.99), p =0.047	1
2 DI	6.32 (3.94-10.13), <0.001	1.18 (0.42-3.29), 0.75	1.26 (0.75-2.11), p = 0.379	1	-
3 DI	2.07 (1.23-3.47), 0.006	1	-	-	-

Table 20. 5-year event-free survival rates and hazard ratios for *ETV6*::*RUNX1* patients stratified by delayed intensification across trials. The latest trial available is used as the baseline in the cox proportional hazards models. DI: delayed intensification.

3.4.3.2 High hyperdiploidy

High hyperdiploidy patients who received two or three delayed intensifications had inferior overall survival when compared to patients who received one DI with 5-year OS rates of 94% and 91% respectively compared to 97% [Figure 33]. Both groups had over twice the risk of death compared to the patients receiving one DI (HR = 2.02, 95% CI (1.35-3.02), p = 0.001 and

HR = 2.64, 95% CI (1.68-4.15), p < 0.001). The difference in survival between 2 and 3 DIs was not significant p = 0.131.



	1 DI	2 DI	3 DI
Hazard Ratio (95% CI), p			
Overall Survival	1	2.02 (1.35-3.02), 0.001	2.64 (1.68-4.15), <0.001

Figure 33. Kaplan Meier and hazard ratios comparing the overall survival of high hyperdiploidy patients stratified by delayed intensification. One delayed intensification is used as the baseline in the Cox proportional hazards models. DI: delayed intensification, OS: overall survival.

When assessing across trials, high hyperdiploidy patients who received 1 delayed intensification on UKALL2003 had significantly better overall survival to those on UKALL2011 with 5-year OS rates of 99% compared to 96% and a hazard ratio of 0.28, 95% CI (0.09-0.84), p = 0.024 [Table 21]. Patients receiving two delayed intensifications had similar 5-year OS

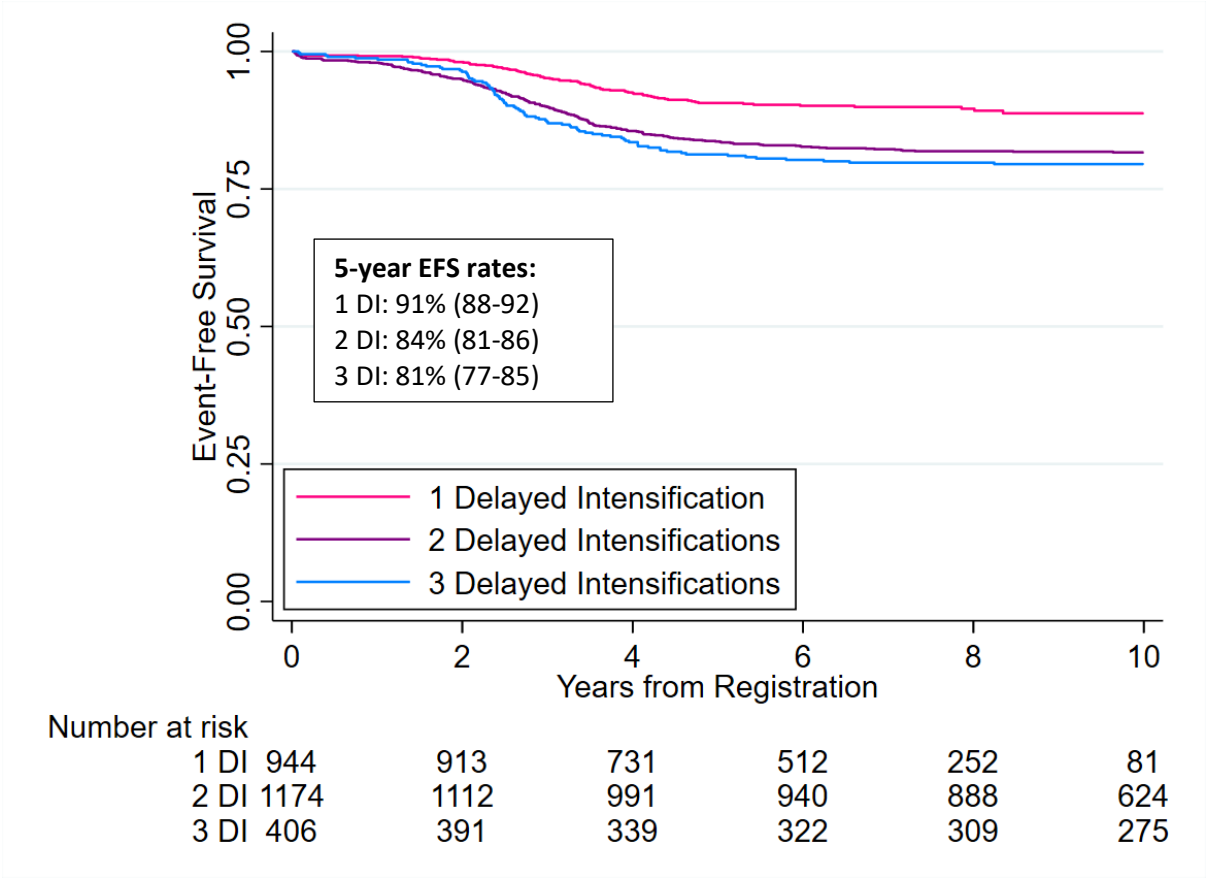
rates on each trial of ~93%, however, there was a significant difference in hazard between UKALLXI92 and UKALL2003 patients at 2.10, 95% CI (1.33-3.31), p = 0.001. The difference between UKALL97 and UKALL2003 was not significant however, despite almost identical 5-year OS rates between UKALLXI92 and UKALL97. There was no significant difference between the patients receiving three delayed intensifications with 5-year OS rates of ~91% and a hazard of 1.28, 95% CI (0.72-2.26), p = 0.405 for UKALLXI92 patients compared to UKALL97 patients. Like in the *ETV6*::*RUNX1* subgroup, high hyperdiploidy patients receiving 1 DI on UKALL2003 also had a higher 5-year OS rate than those receiving 2 DIs (99% v 94%) which was significant with a hazard ratio of 2.76, 95% CI (1.18-6.47), p = 0.019. There was no difference in OS between the two and three delayed intensification groups, both on UKALLXI92 and UKALL97 (p = 0.541 and p = 0.891 respectively).

	UKALLXI92	**UKALL97**	**UKALL97/99**	**UKALL2003**	**UKALL2011**
5-year overall survival rates (95% CI)					
1 DI	-	-	-	99% (97-99.8)	96% (95-98)
2 DI	91% (86-94)	92% (81-97)	95% (92-97)	94% (92-96)	-
3 DI	90% (85-94)	92% (88-95)	-	-	-
Overall survival hazard ratio (95% CI), p-value					
1 DI	-	-	-	0.28 (0.09-0.84), p =0.024	1
2 DI	2.10 (1.33-3.31), 0.001	1.33 (0.57-3.11), 0.513	1.04 (0.61-1.78), p = 0.887	1	-
3 DI	1.28 (0.72-2.26), 0.405	1	-	-	-

Table 21. 5-year overall survival rates and hazard ratios for high hyperdiploidy patients stratified by delayed intensification across trials. The latest trial available is used as the baseline in the Cox proportional hazards models. DI: delayed intensification.

In event-free survival, there was a difference in survival between the delayed intensifications in the high hyperdiploidy subgroup (5-year EFS rates: 91% vs 84% vs 81) as shown in Figure 34. The difference was significant between two delayed intensifications (log-rank p-value <

0.001) and three delayed intensifications (HR: 2.13 (95% CI (1.58-2.86), p < 0.001)) as compared to one delayed intensification. There was no significant difference in survival between two and three delayed intensifications (p = 0.239).



	1 DI	2 DI	3 DI
Hazard Ratio (95% CI), p			
Event-Free Survival	1	Log-rank p-value <0.001	2.13 (1.58-2.86), <0.001

Figure 34. Kaplan Meier and hazard ratios comparing the event-free survival of high hyperdiploidy patients stratified by delayed intensification. One delayed intensification is used as the baseline in the Cox proportional hazards models. The log-rank p-value is given in instances where the proportional hazards assumption was violated. DI: delayed intensification, EFS: event-free survival.

Across trials, patients receiving 2 delayed intensifications had comparable EFS rates with the exception of UKALLXI92 patients who had a 5-year EFS rate of 64% compared to 84%, 85%, and 90% on the three latter trials [Table 22]. Patients receiving 1 delayed intensification on

UKALL2003 had superior event-free survival with a 5-year EFS rate of 96% compared to 89% for patients on UKALL2011. The hazard for UKALL2003 patients was significantly reduced at 0.46, 95% CI (0.26-0.82), p = 0.008. There was no difference in outcome for patients receiving 3 DIs on UKALLXI92 and UKALL97 with a hazard ratio of 1.39, 95% CI (0.91-2.12), p = 0.130 for the UKALLXI92 group. Once again, patients had superior survival when receiving 1 DI on UKALL2003 compared to 2 DIs, with 5-year EFS rates of 96% and 90% respectively (p = 0.028).

	UKALLXI92	**UKALL97**	**UKALL97/99**	**UKALL2003**	**UKALL2011**
5-year event-free survival rates (95% CI)					
1 DI	-	-	-	96% (93-98)	89% (86-91)
2 DI	64% (57-70)	84% (72-91)	85% (80-88)	90% (87-92)	-
3 DI	77% (70-82)	85% (79-89)	-	-	-
Event-free survival hazard ratio (95% CI), p-value					
1 DI	-	-	-	0.46 (0.26-0.82), p =0.008	1
2 DI	4.06 (2.97-5.55), <0.001	1.70 (0.93-3.13), p = 0.086	1.44 (0.99-2.08), p = 0.057	1	-
3 DI	1.39 (0.91-2.12), 0.130	1	-	-	-

Table 22. 5-year event-free survival rates and hazard ratios for high hyperdiploidy patients stratified by delayed intensification across trials. The latest trial available is used as the baseline in the Cox proportional hazards models. DI: delayed intensification.

3.4.3.3 Representative cohort analysis

Again, the cohorts were assessed across the groups to ensure these findings were not caused by an enrichment of patients with certain demographics in one of the delayed intensification subgroups. Table 23 shows there is a higher proportion of males receiving two delayed intensifications within *ETV6*::*RUNX1* patients, however this is likely due to the fact that there was a higher proportion of males than females on UKALL97/99 and patients on this trial all received two delayed intensifications. There is no difference in age or white cell count across the three delayed intensification groups with a median age of ~4, and ~83% of patients having

a low white cell count in each group. A slightly higher proportion of patients receiving two delayed intensifications were treated on regimen A when compared to one delayed intensification whilst the inverse is seen for regimen B, which has a higher proportion in the one delayed intensification group. The proportion of patients who received regimen C was similar at 14% and 12% for 1 and 2 delayed intensifications respectively.

	Total	1 DI	2 DI	3 DI	P-value
Total	1750 (100)	779 (45)	766 (44)	205 (12)	
Median Follow-up (years)	9.62	7.12	11.22	12.16	
Sex					
Male	955 (55)	403 (52)	445 (58)	107 (52)	
Female	795 (45)	376 (48)	321 (42)	98 (48)	p = 0.033
Age (years)					
Median	4	4	4.13	4.27	
1-4	1125 (64)	500 (64)	495 (65)	130 (63)	
5-9	496 (28)	218 (28)	214 (28)	64 (31)	
10-14	111 (6)	50 (6)	50 (7)	11 (5)	
15-19	17 (1)	10 (1)	7 (1)	0 (0)	
≥20	1 (0.06)	1 (0.13)	0 (0)	0 (0)	p = 0.745
White Cell Count ($\times 10^9$/L)					
Median	10.2	9.4	10.15	14.2	
<50	1462 (84)	653 (84)	639 (83)	170 (83)	
≥50	288 (16)	126 (16)	127 (17)	35 (17)	p = 0.946
Regimen					
A	981 (68)	506 (65)	475 (71)	-	
B	274 (19)	165 (21)	109 (16)	-	
C	189 (13)	108 (14)	81 (12)	-	p = 0.026

Table 23. Distribution of *ETV6*::*RUNX1* cases across delayed intensifications by key demographic features. P-values calculated using Pearson χ^2 statistic. DI: delayed intensification.

There was no significant difference in the proportions of males and females, nor the white cell counts across the DI groups within the high hyperdiploidy subgroup. The proportion of patients under the age of 5 is almost identical across the groups at ~63%, however there is a significant difference in the proportion of patients in the subgroups above this age as seen in Table 24. There were proportionally more low risk patients who received two DIs compared

to one DI (67% vs 59%) whilst more patients who received 1 DI were transferred to regimen C (26% vs 20%). However, only patients on UKALL2011 were able to receive one delayed intensification on regimen C, thus this could explain the disparity between the two groups.

	Total	1 DI	2 DI	3 DI	p-value
Total	2524 (100)	944 (37)	1174 (47)	406 (16)	
Median Follow-up (years)	9.39	6.42	11.27	12.18	
Sex					
Male	1393 (55)	530 (56)	636 (54)	227 (56)	
Female	1131 (45)	414 (44)	538 (46)	179 (44)	P = 0.630
Age (years)					
Median	3.95	3.65	4.02	3.89	
1-4	1599 (63)	604 (64)	738 (63)	257 (63)	
5-9	611 (24)	216 (23)	282 (24)	113 (28)	
10-14	212 (8)	69 (7)	113 (10)	30 (7)	
15-19	92 (4)	47 (5)	39 (3)	6 (1)	
≥20	10 (0.4)	8 (1)	2 (0.17)	0 (0)	p = 0.002
White Cell Count ($\times 10^9$/L)					
Median	7.2	7	7.3	7.65	
<50	2297 (91)	866 (92)	1070 (91)	361 (89)	
≥50	227 (9)	78 (8)	104 (9)	45 (11)	p = 0.245
Regimen					
A	1166 (63)	560 (59)	606 (67)	-	
B	260 (14)	143 (15)	117 (13)	-	
C	427 (23)	241 (26)	186 (20)	-	p = 0.004

Table 24. Distribution of high hyperdiploidy cases across delayed intensifications by key demographic, clinical and treatment features. P-values calculated using Pearson χ^2 statistic. DI: delayed intensification.

3.4.4 Comparison of two vs three intensification blocks in UKALLXI92 and UKALL97

3.4.4.1 ETV6::RUNX1

In order to determine if the inferior survival seen in the *ETV6*::*RUNX1* patients who received a third block when compared to just two intensification blocks is legitimate, a comparison of these two groups in just the trials where a randomisation between them occurred was performed. To determine if any results seen were unique to *ETV6*::*RUNX1*, the same analysis was performed in high hyperdiploidy.

Figure 35 shows that for *ETV6*::*RUNX1* patients, there is no significant difference in outcome between the two randomised groups on UKALLXI92 with 5-year OS rates that were identical at 89%. The hazard ratio was not significant at 1.29, 95% CI (0.56-2.94), p = 0.549 [Table 25]. A hazard ratio was unavailable due to a lack of events in the two delayed randomisations group for patients on UKALL97 and is thus not shown in Table 25. Patients on UKALL97 receiving a third intensification block had inferior overall survival compared to patients receiving two intensification blocks. At 5 years, there had been no deaths in the group who received two intensifications whilst the survival for the third intensification block group was at 88% (Log-rank test p-value = 0.0262).

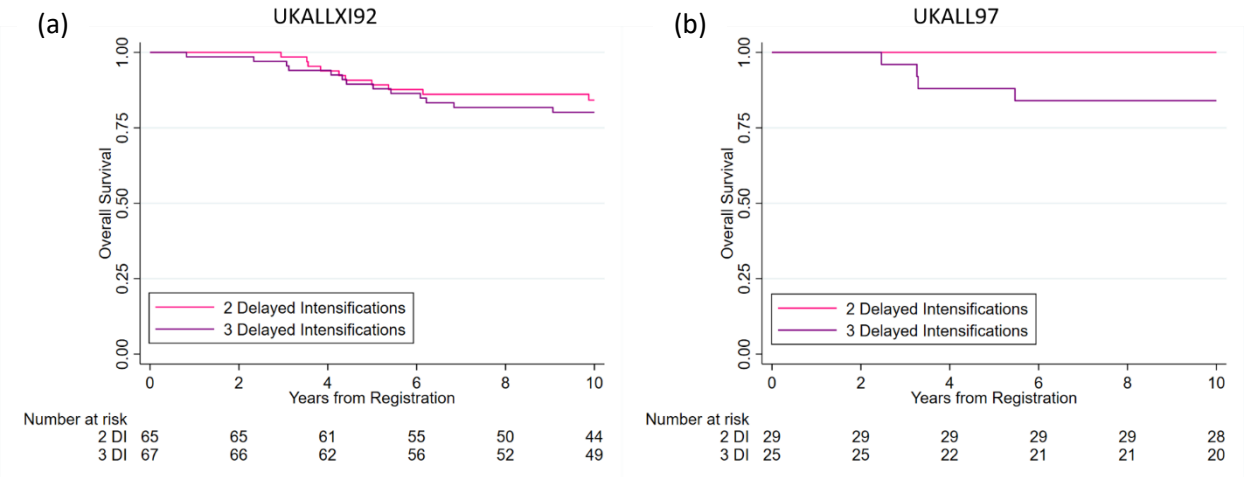


Figure 35. Kaplan Meier depicting the overall survival of *ETV6*::*RUNX1* patients stratified by 3rd delayed intensification block randomisation on (a) UKALLXI92 and (b) UKALL97.

	3rd Block	Not
Hazard Ratio (95% CI), p		
Overall Survival	1.29 (0.56-2.94), 0.549	1

Table 25. Hazard ratio in the overall survival setting of *ETV6*::*RUNX1* cases randomised between two and three delayed intensifications on UKALLXI92. Two delayed intensifications was used as the baseline in the Cox proportional hazards models.

In the event-free survival setting, there was no significant difference in outcome between the groups on UKALLXI92 with 5-year EFS rates of 64% and 57% for 1 and 2 delayed intensifications respectively (p = 0.546). Within UKALL97, there was a significant difference in survival by log-rank test with EFS rates of 97% vs 76% (p = 0.0405) for 2 DI vs 3 DI respectively. Whilst the log-rank test determined there was a difference in the number of observed vs expected events in these groups, there was no significant difference in the risk of an event by the Cox proportional hazards model with p-values of 0.056 [Table 26]. Kaplan-Meiers show these differences in Figure 36.

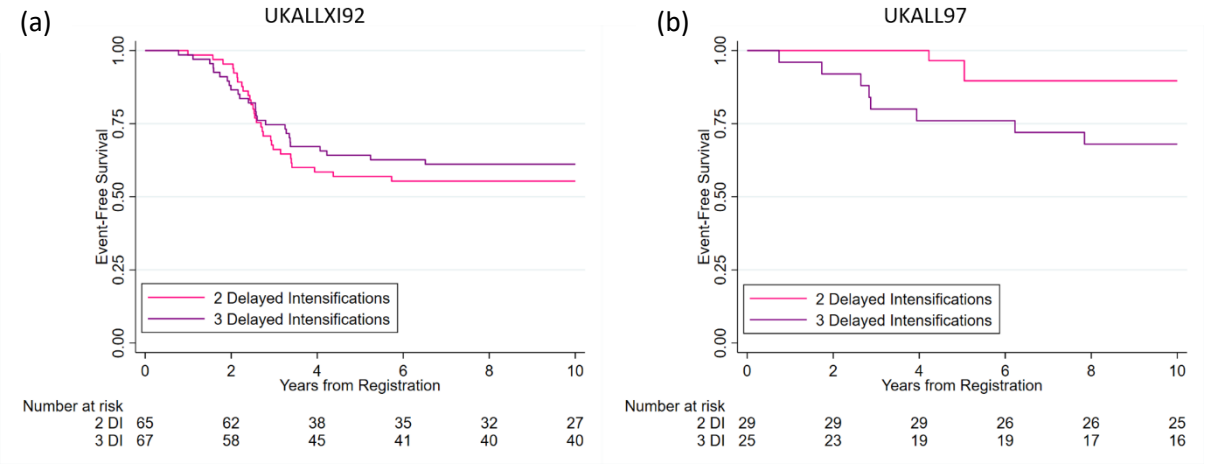


Figure 36. Kaplan Meier depicting the event-free survival of *ETV6*::*RUNX1* patients stratified by 3rd delayed intensification block randomisation on (a) UKALLXI92 and (b) UKALL97.

(a)

	3rd Block	Not
Hazard Ratio (95% CI), p		
Event-Free Survival	0.85 (0.50-1.44), 0.546	1

(b)

	3rd Block	Not
Hazard Ratio (95% CI), p		
Event-Free Survival	3.65 (0.97-13.75), 0.056	1

Table 26. Hazard ratio in the event-free survival setting of *ETV6*::*RUNX1* cases randomised between two and three delayed intensifications on (a) UKALLXI92 and (b) UKALL97. Two delayed intensifications was used as the baseline in the Cox proportional hazards models.

3.4.4.2 High hyperdiploidy

It is clear from Figure 37 that high hyperdiploidy patients had identical overall survival rates at 5 years at 90% (95% CI (85-94)). The hazard was non-significant (p = 0.572). In UKALL97, patients receiving three DIs had poorer outcomes compared to the other group (5-year OS rates: 89% vs 94%), but that difference wasn't significant with a Cox proportional hazards model p-value of 0.427 [Table 27].

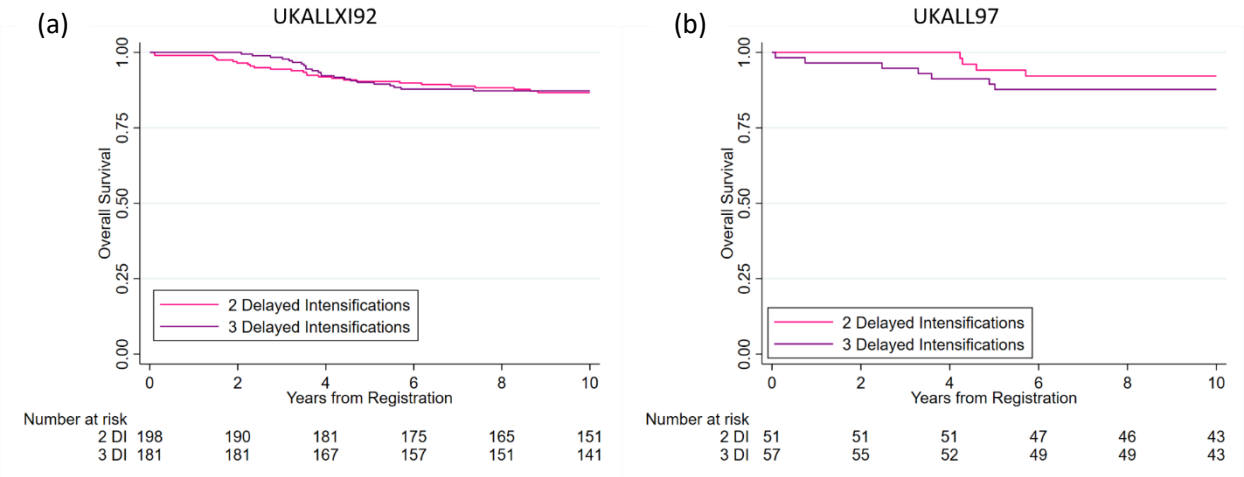


Figure 37. Kaplan Meier depicting the overall survival of high hyperdiploidy patients stratified by 3rd delayed intensification block randomisation on (a) UKALLXI92 and (b) UKALL97.

(a)

	3rd Block	Not
Hazard Ratio (95% CI), p		
Overall Survival	0.86 (0.50-1.47), 0.572	1

(b)

	3rd Block	Not
Hazard Ratio (95% CI), p		
Overall Survival	1.65 (0.48-5.62), 0.427	1

Table 27. Hazard ratio in the overall survival setting of high hyperdiploidy cases randomised between two and three delayed intensifications on (a) UKALLXI92 and (b) UKALL97. Two delayed intensifications was used as the baseline in the Cox proportional hazards models.

High hyperdiploidy patients who received a third intensification block on UKALLXI92 had significantly better event-free survival than those who received two DIs with 5-year EFS rates of 77% vs 63% and a hazard ratio of 0.54 (95% CI (0.37-0.77), p = 0.001) [Figure 38, Table 28]. This is in contrast to the results seen in Section 3.4.7 in which third block patients had the worst survival. Interestingly, in UKALL97, the hazard ratio suggests an increased risk in the three intensification blocks group (HR: 1.51, 95% CI (0.65-3.48), p = 0.338), which is the inverse of what was seen in UKALLXI92. However, this risk is not significant with the 95% confidence interval spanning hazards both above and below the baseline of 1.

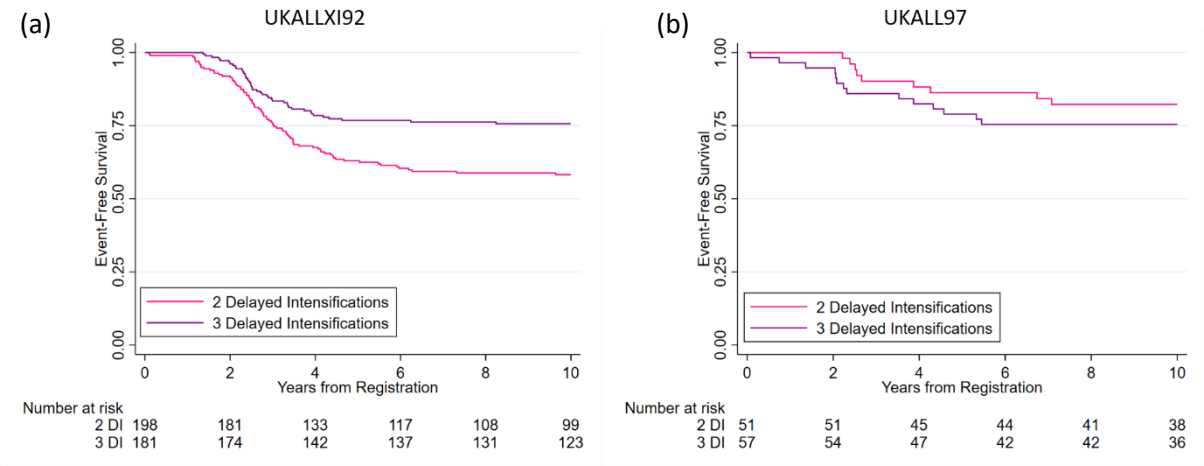


Figure 38. Kaplan Meier depicting the event-free survival of high hyperdiploidy patients stratified by 3rd delayed intensification block randomisation on (a) UKALLXI92 and (b) UKALL97.

(a)

	3rd Block	Not
Hazard Ratio (95% CI), p		
Event-Free Survival	0.54 (0.37-0.77), 0.001	1

(b)

	3rd Block	Not
Hazard Ratio (95% CI), p		
Event-Free Survival	1.50 (0.65-3.47), 0.340	1

Table 28. Hazard ratio in the event-free survival setting of high hyperdiploidy cases randomised between two and three delayed intensifications on (a) UKALLXI92 and (b) UKALL97. Two delayed intensifications was used as the baseline in the Cox proportional hazards models.

3.4.4.3 Representative cohort analysis

Again, the demographics across the two groups were compared within each genetic subgroup. Event-free survival switched between being superior in the patients receiving a third intensification block treated on UKALLXI92 to the patients only receiving two intensification blocks when treated on UKALL97, however this same phenomenon was not seen in overall survival. Therefore, it was hypothesised that this change in EFS between the trials was due to a reduction in the number of relapses in patients treated with two delayed intensifications between UKALLXI92 and UKALL97. Thus, relapse site was included for

comparison in this analysis to determine if a specific site of relapse (e.g. CNS) was the cause of this finding.

For the *ETV6*::*RUNX1* subgroup, within UKALLXI92 there was no difference in the proportions of sex, age, white cell count, or CNS treatment which is expected as this was accounted for in the randomisation [Table 29]. There was also no significant difference in the proportion of relapses or deaths between the two groups. Relapse site also showed no significant difference with approximately half of all relapses being in the bone marrow. Thus, any difference in outcome is likely due to the difference in treatment alone.

UKALLXI92	Total	3rd Block	No 3rd Block	p-value
Total, n(%)	132 (100)	67 (51)	65 (49)	
Sex				
Male	69 (52)	35 (52)	34 (52)	
Female	63 (48)	32 (48)	31 (48)	p = 0.994
Age (years)				
1-4	82 (62)	42 (63)	40 (62)	
5-9	38 (29)	19 (28)	19 (29)	
10-14	12 (9)	6 (9)	6 (9)	
15-19	0 (0)	0 (0)	0 (0)	
≥20	0 (0)	0 (0)	0 (0)	p = 0.991
White Cell Count ($\times 10^9$/L)				
<50	108 (82)	55 (82)	53 (82)	
>50	24 (18)	12 (18)	12 (18)	p = 0.935
CNS treat				
IT-MTX	58 (44)	26 (39)	32 (49)	
HD-MTX	64 (48)	38 (57)	26 (40)	
Radiotherapy	10 (8)	3 (4)	7 (11)	p = 0.109
Relapse				
Yes	55 (42)	26 (39)	29 (45)	
No	77 (58)	41 (61)	36 (55)	p = 0.499
Relapse Site				
Bone Marrow	26 (47)	13 (50)	13 (45)	
CNS	8 (15)	5 (19)	3 (10)	
BM + CNS	9 (16)	3 (12)	6 (21)	
Testes	6 (11)	2 (8)	4 (14)	
Other	6 (11)	3 (12)	3 (10)	p = 0.734
Dead				
Yes	23 (17)	13 (19)	10 (15)	
No	109 (83)	54 (81)	55 (85)	p = 0.543

Table 29. Distribution of *ETV6*::*RUNX1* cases randomised to two or three delayed intensification blocks on UKALLXI92 by key demographic, clinical and treatment features. P-values calculated using Pearson χ^2 statistic.

In UKALL97, there was also no significant difference in the proportions of sex, age, white cell count, or CNS treatment between the randomisations. The HR1 group, comprised of patients who were treated on a regimen for children with high-risk acute lymphoblastic leukaemia [Section 2.2.3], is driving the difference in white cell count and CNS treatment seen in Table 30, however this isn't detrimental to the comparison performed in this section as the HR1 was excluded due to differences in treatment. There was also no difference in the proportion of relapses or deaths between these groups, nor the type of relapse. There were only 3 relapses in the group receiving two intensifications and they were all marrow relapses, whilst in the three intensifications group, 5 out of 8 relapses were marrow. There are no isolated CNS relapses in either group which accounted for 10% and 19% of all relapses in the two DI and three DI groups in UKALLXI92 respectively. This could imply that the CNS treatment administered on the UKALL97 trial was more effective than that administered in UKALLXI92. The difference in the CNS treatment administered between the two trials was that patients treated on UKALL97 with a white cell count $<50x10^9$/L were no longer randomised between IT-MTX or HD-MTX (as in UKALLXI92) and were instead all assigned IT-MTX. Therefore, the number of patients who received IT-MTX on UKALL97 approximately doubled, whilst the number who received HD-MTX decreased dramatically [Tables 29 and 30]. This difference may have improved EFS in the group receiving two intensification blocks, removing the need for three intensifications in the *ETV6*::*RUNX1* subgroup.

UKALL97	Total	3rd Block	No 3rd Block	HR1	p-value
Total, n(%)	56 (100)	25 (45)	29 (52)	2 (3)	
Sex					
Male	29 (52)	12 (48)	15 (52)	2 (100)	
Female	27 (48)	13 (52)	14 (48)	0 (0)	p = 0.367
Age (years)					
1-4	33 (59)	16 (64)	17 (59)	0 (0)	
5-9	20 (36)	8 (32)	10 (34)	2 (100)	
10-14	3 (5)	1 (4)	2 (7)	0 (0)	
15-19	0 (0)	0 (0)	0 (0)	0 (0)	
≥20	0 (0)	0 (0)	0 (0)	0 (0)	p = 0.402
White Cell Count ($\times 10^9$/L)					
<50	42 (75)	20 (80)	22 (76)	0 (0)	
>50	14 (25)	5 (20)	7 (24)	2 (100)	p = 0.042
CNS treat					
IT-MTX	42 (75)	20 (80)	22 (76)	0 (0)	
HD-MTX	8 (14)	2 (8)	4 (14)	2 (100)	
Radiotherapy	6 (11)	3 (12)	3 (10)	0 (0)	p = 0.012
Relapse					
Yes	11 (20)	8 (32)	3 (10)	0 (0)	
No	45 (80)	17 (68)	26 (90)	2 (100)	p = 0.106
Relapse Site					
Bone Marrow	8 (73)	5 (63)	3 (100)	0 (0)	
CNS	0 (0)	0 (0)	0 (0)	0 (0)	
BM + CNS	2 (18)	2 (25)	0 (0)	0 (0)	
Testes	1 (9)	1 (13)	0 (0)	0 (0)	
Other	0 (0)	0 (0)	0 (0)	0 (0)	p = 0.461
Dead					
Yes	4 (7)	4 (16)	0 (0)	0 (0)	
No	52 (93)	21 (84)	29 (100)	2 (100)	p = 0.069

Table 30. Distribution of *ETV6*::*RUNX1* cases randomised to two or three delayed intensification blocks on UKALL97 by key demographic, clinical and treatment features. P-values calculated using Pearson χ^2 statistic.

Within high hyperdiploidy, there was once again no difference in the proportions of age, sex, white cell count, or CNS treatment between the randomised groups on UKALLXI92. There was however a difference in the proportion of relapses between the group, with the patients

receiving two delayed intensifications having proportionally more relapses 38% vs 25% (p = 0.007). There is no difference in the type of relapse suggesting that this difference is not driven by a particular type of relapse. There is no difference in the proportion of deaths across the groups suggesting that the additional relapses in the two delayed intensification block group were salvageable [Table 31].

	Total	3rd Block	No 3rd Block	p-value
Total, n(%)	379 (100)	181 (48)	198 (52)	
Sex				
Male	220 (58)	109 (60)	111 (56)	
Female	159 (42)	72 (40)	87 (44)	p = 0.412
Age (years)				
1-4	237 (63)	113 (62)	124 (63)	
5-9	111 (29)	54 (30)	57 (29)	
10-14	31 (8)	14 (8)	17 (9)	
15-19	0 (0)	0 (0)	0 (0)	
≥20	0 (0)	0 (0)	0 (0)	p = 0.942
White Cell Count ($\times 10^9$/L)				
<50	348 (92)	165 (91)	183 (92)	
>50	31 (8)	16 (9)	15 (8)	p = 0.654
CNS treat				
IT-MTX	181 (48)	91 (51)	90 (46)	
HD-MTX	184 (49)	81 (45)	103 (52)	
Radiotherapy	11 (3)	7 (4)	4 (2)	p = 0.273
Relapse				
Yes	120 (32)	45 (25)	75 (38)	
No	259 (68)	136 (75)	123 (62)	P = 0.007
Relapse Site				
Bone Marrow	40 (41)	19 (42)	30 (40)	
CNS	17 (14)	7 (16)	10 (13)	
BM + CNS	20 (17)	8 (18)	12 (16)	
Testes	11 (9)	4 (9)	7 (9)	
Other	23 (19)	7 (16)	16 (21)	p = 0.954
Dead				
Yes	54 (14)	24 (13)	30 (15)	
No	325 (86)	157 (87)	168 (85)	p = 0.599

Table 31. Distribution of high hyperdiploidy cases randomised to two or three delayed intensification blocks on UKALLXI92 by key demographic, clinical and treatment features. P-values calculated using Pearson χ^2 statistic.

Within UKALL97, there was a significant difference in sex, age and CNS treatment, with a higher proportion of females receiving two intensification blocks as well as older patients overall, and more patients receiving radiotherapy in this subgroup compared to those

receiving three delayed intensifications. There was however no significant difference in the proportion of relapses or deaths in those groups, and the site of relapse remained approximately stable between the groups also. This is seen in Table 32 below.

	Total	3rd Block	No 3rd Block	HR1	p-value
Total, n(%)	122 (100)	57 (47)	51 (42)	14 (11)	
Sex					
Male	69 (57)	35 (61)	22 (43)	12 (86)	
Female	53 (43)	22 (39)	29 (57)	2 (14)	p = 0.010
Age (years)					
1-4	70 (57)	36 (63)	34 (67)	0 (0)	
5-9	34 (28)	19 (33)	11 (22)	4 (29)	
10-14	12 (10)	2 (4)	5 (10)	5 (36)	
15-19	6 (5)	0 (0)	1 (2)	5 (36)	
≥20	0 (0)	0 (0)	0 (0)	0 (0)	p < 0.001
White Cell Count ($\times 10^9$/L)					
<50	107 (88)	51 (89)	46 (90)	10 (71)	
>50	15 (12)	6 (11)	5 (10)	4 (29)	p = 0.142
CNS treat					
IT-MTX	94 (79)	51 (91)	43 (88)	0 (0)	
HD-MTX	23 (19)	5 (9)	4 (8)	14 (100)	
Radiotherapy	2 (2)	0 (0)	2 (4)	0 (0)	p < 0.001
Relapse					
Yes	23 (19)	12 (21)	9 (18)	2 (14)	
No	99 (81)	45 (79)	42 (82)	12 (86)	p = 0.811
Relapse Site					
Bone Marrow	12 (52)	6 (50)	5 (56)	1 (50)	
CNS	3 (13)	2 (17)	1 (11)	0 (0)	
BM + CNS	5 (22)	3 (25)	2 (22)	0 (0)	
Testes	0 (0)	0 (0)	0 (0)	0 (0)	
Other	3 (13)	1 (8)	1 (11)	1 (50)	p = 0.776
Dead					
Yes	14 (11)	7 (12)	4 (8)	3 (21)	
No	108 (89)	50 (88)	47 (92)	11 (79)	p = 0.356

Table 32. Distribution of high hyperdiploidy cases randomised to two or three delayed intensification blocks on UKALL97 by key demographic, clinical and treatment features. P-values calculated using Pearson χ^2 statistic.

There is a greater reduction in the proportion of relapses from UKALLXI92 to UKALL97 in the patients receiving two delayed intensifications compared to those receiving three delayed intensifications (*ETV6*::*RUNX1*: 39% vs 45% in UKALLXI92 and 32% vs 10% in UKALL97, high hyperdiploidy: 25% vs 38% in UKALLXI92 and 21% vs 18% in UKALL97) which could be driving the difference in the hazard ratios seen in the two trials [Tables 29-32]. It is clear from Figures 35 - 38 that the risk of an event for patients receiving two intensification blocks drastically reduced between the two trials. This same difference is not seen between the two third intensification block groups. Furthermore, whilst not statistically significant, the Cox proportional hazards model suggests there is a reduced risk of an event for patients receiving a third block compared to two delayed intensifications in UKALLXI92, whilst in UKALL97, it is suggested that a third intensification block increases the risk of an event. Thus, a test of heterogeneity between the third intensifications blocks on the two trials was performed within *ETV6*::*RUNX1* [Figure 39] and high hyperdiploidy subgroups [Figure 40] within the relapse setting – the primary driver of the difference seen in the event-free setting. These tests determined that there was evidence of heterogeneity between the randomised groups in *ETV6*::*RUNX1* patients ($p < 0.01$) but no heterogeneity in the high hyperdiploidy subgroup due to a p-value > 0.05.

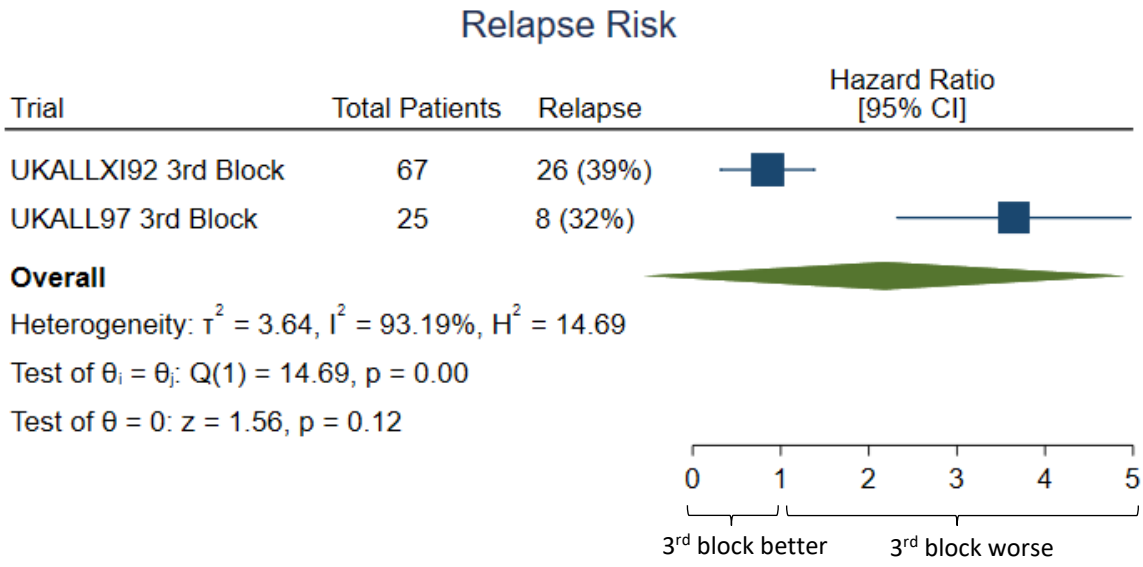


Figure 39. Forest plot and test of heterogeneity comparing risk of relapse between *ETV6*::*RUNX1* patients receiving three delayed intensifications on UKALLXI92 and UKALL97.

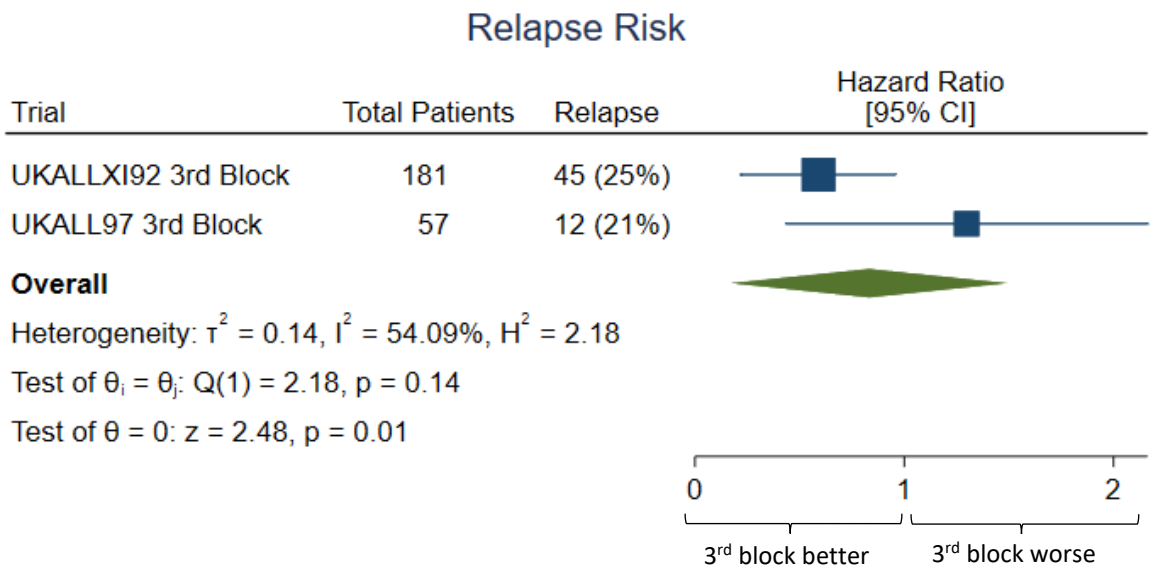


Figure 40. Forest plot and test of heterogeneity comparing risk of relapse between high hyperdiploidy patients receiving three delayed intensifications on UKALLXI92 and UKALL97.

3.5 Discussion

This chapter ascertains the survival of good risk genetics patients on historic UKALL paediatric trials and how different treatment elements affect these survival rates. This study also demonstrated that *ETV6*::*RUNX1* and high hyperdiploidy patients were enriched with good risk features including a younger age and lower white cell counts which is consistent with the literature (Zheng *et al.*, 2021; Alvarez *et al.*, 2007; Paulsson and Johansson, 2009; Haas and Borkhardt, 2022).

There is improvement in survival in both endpoints from the first trial considered in this study to the latest trial in both subgroups, with each trial generally improving upon the last with few exceptions [Figures 23-26]. Within *ETV6*::*RUNX1* there is a 7% increase in overall survival across the trials at 5 years and a 30% increase in event-free survival from UKALLXI92 to UKALL2003 which had the EFS rates at 94% [Figures 23 and 24]. For high hyperdiploidy, a 5% increase in overall survival and a 21% increase in event-free survival was achieved at the same time point [Figures 25 and 26]. In *ETV6*::*RUNX1*, cure rates were comparable from UKALL97 whilst event-free survival rates surpassed 90% from UKALL97/99, suggesting that the change in treatment stratification to regimen based on NCI risk was beneficial but that no changes to treatment in the latter trials significantly affected survival. These findings support those of Østergaard *et al.* who concluded that the outcomes between contemporary trials (enrolling

between 2000-2017) were similar and that further treatment de-escalation should be assessed for feasibility in the *ETV6*::*RUNX1* population (Østergaard *et al.*, 2024). This same effect was seen within overall survival in the high hyperdiploidy subgroup (with the earlier two trials having significantly worse cure rates) but was not the case in event-free survival with UKALL97/99 having similar rates to UKALL97 and the latter two trials having significantly better EFS rates ($p < 0.001$) [Figures 25 and 26].

As a whole, and within both genetic subgroups considered in this study, UKALL2011 had inferior event-free survival rates to its predecessor UKALL2003, but this does not impact overall survival in either subgroup. This engenders a delicate question concerning the optimal approach to treatment: should we give less intensive frontline therapy to all and salvage relapses (particularly in low risk subgroups) or should we continue with more intensive frontline therapy, reducing relapses but potentially causing unnecessary toxicities and long-term late effects in patients? Further investigation into different treatment elements, drug dosages, novel therapies, and pharmacokinetics may be required to elucidate on this matter; helping to determine how many patients would need to be potentially over-treated in order to prevent these events and if there are any additional factors that could be used to further stratify patients.

A limitation of this analysis is that patients were assigned to regimen based on the risk factors age, white cell count at diagnosis, and MRD as outlined in Chapter 2. Therefore, it is expected that the high-risk patients assigned to regimen C would perform poorly compared to those on regimens A and B. In order to mitigate this limitation, a representative cohort analysis was performed to examine the distribution of these risk factors. When comparing the distribution of cases across the trials by key features, it became evident that more patients from each subgroup were classified as high risk over time and moved to regimen C, which raised the question as to whether this change was beneficial for these patients. For *ETV6*::*RUNX1* patients in this study, patients treated on regimen B and C have inferior overall survival rates to those treated on regimen A, which is a result not seen in high hyperdiploidy patients who have similar 5 year survival rates and non-significant p-values in Cox proportional hazards models [Figures 27 and 29]. This suggests either that NCI risk/ early treatment response are still prognostic within the *ETV6*::*RUNX1* subgroup but not within high hyperdiploidy, or that the more intensive regimens have an adverse effect on the treatment of *ETV6*::*RUNX1*

patients. Further to this, the significant difference in event-free survival between regimens A and C in both subgroups supports the hypothesis that slow early response is prognostic of events [Figures 28 and 30]. Also of note, is that good risk genetics patients could only transfer to Regimen C due to slow early response, including EOI MRD from UKALL2003 onwards. Hence, due to the increase of this phenomena over the course of the trials, one could hypothesise that the overall reduction in intensity of induction resulted in fewer good risk genetics patients responding in a timely manner, resulting in more intensive treatment for these patients for the rest of the protocol. However, further research into this matter is required.

Initially, comparison of outcomes by the number of delayed intensifications received as a whole was performed. Within both subgroups, each increase in the number of delayed intensifications resulted in inferior survival rates and a significantly larger likelihood of an event occurring as specified by the Cox proportional hazards models. Analysis across trials generally confirmed this finding with patients on UKALL97 having equivalent or superior survival when receiving 2 DIs compared to 3, and patients on UKALL2003 having better survival rates when receiving 1 DI compared to 2. However, this difference for UKALL2003 patients is likely because the 2 DI group was enriched for regimen C patients as they were assigned to 2 delayed intensifications. Thus, further multivariate analysis between treatment elements is required. The results between the 2 and 3 DI groups treated on UKALL97 were of complete contrast to those of UKALLXI92, wherein having 3 delayed intensifications generally improved survival. Whilst this result was interesting, it was possible that this observation was skewed by patients assigned to a certain number of DIs on a trial. Thus, in order to determine the validity of these findings, analysis was performed comparing the patients randomised to receive two delayed intensifications vs three delayed intensifications on UKALLXI2 and UKALL97.

Within both subtypes, there was a much larger reduction in proportion of relapses in the group randomised to receive two delayed intensifications from UKALLXI92 to UKALL97 than those receiving three intensifications [Tables 29-32]. For example, within *ETV6*::*RUNX1*, the reduction in the two DI group was from 45% to 10%, whilst in the three delayed intensification group it was from 39% to 32%. Furthermore, this effect often resulted in the group receiving two delayed intensifications initially having a proportionally greater number of relapses to

proportionally fewer compared to the third intensification block group. Overall, this reduction in relapses does not seem to be due to a specific type of relapse as within each subtype, the difference in the proportions of relapse site remains statistically non-significant as shown in Table 33. Similarly, in death the same phenomenon occurred in the proportion of deaths between the two trials, with the proportion of deaths remaining relatively stable between the trials within the third intensification block but reducing in the other group. Therefore, whilst the p-values were largely non-significant, the group with the increased risk of an event switched from two delayed intensifications to three delayed intensifications between UKALLXI92 and UKALL97. A test of heterogeneity performed in each subgroup determined that this difference was significant in the *ETV6*::*RUNX1* which suggests that whilst the third block was beneficial for these patients on UKALLXI92, it was detrimental to those patients on UKALL97. Overall, these findings suggest that the additional changes made to the protocol between UKALLXI92 and UKALL97 removed the necessity for a third intensification block in all good risk patients. These findings differ from those of Hann *et al.* who concluded that the results from UKALLXI92 and UKALL97 support the idea that intensification of treatment is beneficial to all children with ALL (Hann *et al.*, 2000). However, as noted in their study, they could not determine if the additional intensification block was the sole cause of the improved outcome in UKALLXI92 and UKALL97 or whether design of the schedule made a contribution. Analysis in this thesis of the two trials separately seems to imply that it was the changes to the treatment schedule on UKALL97, and not the additional intensification block, that caused the improvement.

Genetic Subgroup	Pearson χ^2 test p-value
ETV6::*RUNX1*	P = 0.785
High hyperdiploidy	P = 0.601

Table 33. Results of Pearson χ^2 tests of the difference in proportions of relapse site between UKALLXI92 and UKALL97 within *ETV6*::*RUNX1* and high hyperdiploidy subgroups.

In summary, these findings suggest that the changes in treatment employed throughout the clinical trials from UKALL97 onwards did not significantly affect the cure rates of the two good risk genetic subgroups. As such, the next logical step is to identify which treatment elements

throughout these trials were the least intensive whilst still maintaining optimal outcomes in order to reduce unnecessary toxicities. This is the overarching aim of chapters 4 and 5.

Chapter 4. Calculation and assembly of drug dosages dataset and development of a dose intensity score

4.1 Introduction

Current research efforts in ALL often assess the effect of changing individual treatment elements within protocols, such as additional intensification blocks or the stratification of patients into regimens, on patient outcomes overall. This approach, whilst useful in determining the benefit of these modifications, fails to capture the intensity of all therapy received and its resulting effect on outcomes as a whole. In order to achieve this, some method to compare intensity of the pathways is necessary. However, this is difficult to achieve as no patient with ALL is treated with monotherapy and the synergy of certain drugs can effect overall intensity not captured simply by dosage (Cheok and Evans, 2006; Bayat Mokhtari *et al.*, 2017). Moreover, each patient will respond to therapy differently due to their unique metabolism and genetics (Wu and Li, 2018; Cheok *et al.*, 2009). Thus, where available, studies will harness the use of pharmacokinetics and pharmacogenetics to optimise efficacy of treatment in individual patients (Cheok *et al.*, 2009).

In many studies however, these type of data were not collected, and thus a true assessment of the intensity of treatment and an individual patient's response to therapy cannot be made. As such, attempts have been made within the literature to approximate the intensity of treatment based solely on the dosage of drugs administered. The first study to present this was in 1984 in which Hryniuk *et al.* calculated dose intensities of cytotoxic drugs cyclophosphamide, methotrexate, and 5-fluorouracil (which he called C, M, and F respectively) relative to a chemotherapy regimen described to successfully treat breast cancer which he called Cooper's regimen (Hryniuk and Bush, 1984). Whilst full details are available in the original paper, a brief summary of the Hryniuk method was as follows: 1. Convert doses of C, M, and F to the standard form of mg/m^2/week, 2. Express the dose intensities as a fraction of the dose intensities in Cooper's original regimen, 3. Calculate the average relative dose intensity of the therapy (Hryniuk and Bush, 1984). This method was applied to compare chemotherapy trials in stage II breast cancer and to compare chemotherapy regimens in ovarian carcinoma (Hryniuk and Levine, 1986; Levin and Hryniuk, 1987). Hryniuk also developed an alternative method for calculating dose intensity named summation dose intensity (Hryniuk, Frei and Wright, 1998).

In more recent years, Lee *et al.* established a modified version of Hryniuk's relative dose intensity model for dose reduction of FOLFIRINOX in pancreatic cancer research which was

validated in a study by Vary *et al.* with limited success. (Lee *et al.*, 2017; Vary *et al.*, 2021). In ALL, treatment intensity for *ETV6*::*RUNX1* patients was successfully calculated by a modification of Hryniuk's relative dose intensity method (Østergaard *et al.*, 2024). These approaches all rely on the assignment of a baseline treatment arm/ regimen to calculate a relative dose intensity for the other regimens which is beneficial for the comparison of trials to one established regimen but not as useful when trying to compare the actual intensity of each pathway, which is the aim of this study. Thus, efforts to develop a novel dose intensity score and use it to identify optimal treatment elements are explored in this chapter.

4.2 Aims

The aims of this chapter are:

- Calculate drug dosages for patients by pathway and randomisations on four paediatric clinical trials.

- Develop a novel dose intensity score from these drug dosages to compare the actual treatment intensity of multi-agent chemotherapy regimens across the trials without the need for a baseline regimen.

- Identify optimal treatment elements for good risk genetic subgroups that ensure minimal treatment intensity whilst maintaining excellent outcomes.

4.3 Methods

4.3.1 Calculation of drug dosages

Individual daily drug dosages prescribed in mg/m^2 (and iu/m^2 for asparaginase preparations) were mapped in a dataframe for each possible treatment pathway on the paediatric clinical trials considered in this thesis.

Cumulative drug dosage by treatment phase (e.g. induction, delayed intensification, maintenance, etc.) was calculated by multiplying the daily doses by the number of administrations specified per treatment phase by the protocol. Dexamethasone and prednisolone were grouped as steroids with dexamethasone doses multiplied by 6.7 to

account for its greater anti-leukaemic effect (Østergaard *et al.*, 2024). The purines: 6-mercaptopurine and 6-thioguanine were grouped at a 1:1 ratio. The different administrations of methotrexate were grouped with intrathecal methotrexate multiplied by 1.7 to account for its paradoxical increased systemic exposure compared with oral administration (Bostrom, Erdmann and Kamen, 2003; Bleyer, Nelson and Kamen, 1997; Kose *et al.*, 2009; Finkelstein *et al.*, 2005; Heilmann *et al.*, 2023). Doses of different asparaginase preparations were converted to pegylated L-asparaginase based on conversions specified in the appendices of the trial protocols. Patients were assumed to receive the dose stated in the protocol unless information about deviations was available. Patients on UKALL97/99, UKALL2003, and UKALL2011 receiving Capizzi interim maintenance on regimen C were scheduled for escalating methotrexate and had their dose increased by 50mg/m^2 to toxicity from a starting dose of 100mg/m^2 for 5 doses resulting in a total possible dose of 300mg/m^2. For calculating the dose intensity scores, patients were randomly fitted along a Normal distribution within this range as information of actual dose received was unavailable, and it was assumed that the majority of patients would achieve at least some escalation, with few reaching the highest possible dose and few remaining at the starting dose. Thus, a Normal distribution would approximately model this effect. Similarly, for UKALLXI92 and UKALL97, maintenance purine doses were escalated by 25% every four weeks if the previous dose was tolerated and was adjusted upwards or downwards throughout therapy based on toxicity. Again, random values following a Normal distribution within the range were assigned to patients for calculating the dose intensity score as actual doses were unavailable. Intrathecal methotrexate dosage was assigned based on three possible age groups: <2, 2, or ≥3 years old. These data were used to calculate both the dose intensity and relative dose intensity scores as outlined in Sections 4.3.2 and 4.3.3

For the area under the curve dose intensity score, the daily drug dosages dataframe was used. The calculation for this is outlined in Section 4.3.4. The same steps as above were taken in regards to the grouping of the data, conversions of asparaginase preparations, and assumption of received dose. However, unlike the calculation of drug dosage by treatment phase, for the drugs which were prescribed in escalating doses, it was assumed the maximum dose (300mg/ m^2) had been achieved for methotrexate and that the standard dose (75mg/m^2) was maintained for purine in all patients. Etoposide was only prescribed for patients on

UKALLXI92 and UKALL97, doxorubicin wasn't prescribed until UKALL97/99, and daunorubicin wasn't given to patients assigned to regimen A on any trial and in these instances, the rows were kept blank in the dataframe to denote this.

Patients who received radiotherapy as their CNS treatment were excluded from the calculation and analysis of dose intensity, as this was performed only in patients who received chemotherapy alone as their treatment. Total therapy ran for 100 weeks on UKALLXI92, 105 weeks on UKALL97, and 2 years or 3 years from the start of interim maintenance 1 for girls or boys respectively on UKALL97/99, UKALL2003, and UKALL2011.

4.3.2 Calculation of the dose intensity score

The dose intensity score was calculated using a modified method of that described by Hryniuk *et al.* for each individual treatment phase and overall treatment as described below and exemplified in Table 34 (Hryniuk and Bush, 1984).

Step 1. The cumulative drug dose by phase for each drug was calculated as described in Section 4.3.1.

Step 2. These cumulative dosages for each drug in the phase were then added together.

Step 3. This was then divided by the number of weeks in the phase.

Step 4. For the dose intensity score at each phase, the above value was divided by the number of drugs given at each phase.

Step 5. To get a dose intensity score for the complete protocol, the phase dose intensity scores were added together.

Initial drug dosages	Step 1	Step 2	Step 3	Step 4	Step 5
Induction Phase (4 weeks)					
Steroid = 10mg/m^2 **Methotrexate = 17mg/m^2** **Vincristine = 1.5mg/m^2**	Steroid = 10*28 Meth = 17*2 Vin = 1.5*2	280 + 34 + 3 = 317	317/4 = 79.25	7.125/3 = 26.42	26.42 + 20.52 = 46.94
Consolidation Phase (8 weeks)					
Steroid = 28mg/m^2 **Methotrexate = 37mg/m^2** **Vincristine = 6mg/m^2**	Steroid = 28*15 Meth = 17*4 Vin = 1.5*3	420 + 68 + 4.5 = 492.5	492.5/8 = 61.563	8.25/3 = 20.52	

Table 34. Simple example of process used to calculate the dose intensity score. Meth: methotrexate, vin: vincristine.

4.3.3 Calculation of the relative dose intensity score

The relative dose intensity score was calculated in the same manner as the dose intensity score in Section 4.3.2 with the addition of a step between steps 1 and 2 (named step 1.5) in which the cumulative drug dose is divided by the smallest cumulative dose of that drug given on any of the treatment pathways. This results in the lowest cumulative dose having a value of 1 and every other cumulative dose being a value relative to that baseline dosage. This process is outlined below.

Step 1. The cumulative drug dose by phase for each drug was calculated as described in Section 4.3.1.

Step 1.5. Divide the cumulative drug dose by the minimum cumulative dose of that drug.

Step 2. These cumulative dosages for each drug in the phase were then added together.

Step 3. This was then divided by the number of weeks in the phase.

Step 4. For the dose intensity score at each phase, the above value was divided by the number of drugs given at each phase.

Step 5. To get a dose intensity score for the complete protocol, the phase dose intensity scores were added together.

4.3.4 Calculation of the area under the curve dose intensity score

The calculation for the area under the curve dose intensity score was adopted from Allgoewer *et al.* which present the calculation of the area under the curve using the trapezoid rule as a way to study participants' trajectories (Allgoewer *et al.*, 2018). Cumulative daily drug dosages were calculated for each pathway on the aforementioned trials from the daily drug dosages table described in Section 4.3.1, and these values were plotted to produce a curve. This curve was then split into n trapezoid shapes, where n is the number of days in each pathway, and the area of those trapezoid shapes is calculated using the formula for the area of a trapezoid. The AUC dose intensity score is then calculated by adding all of the trapezoid shape areas together to get a total area of the whole curve. This process is illustrated in Figure 41.

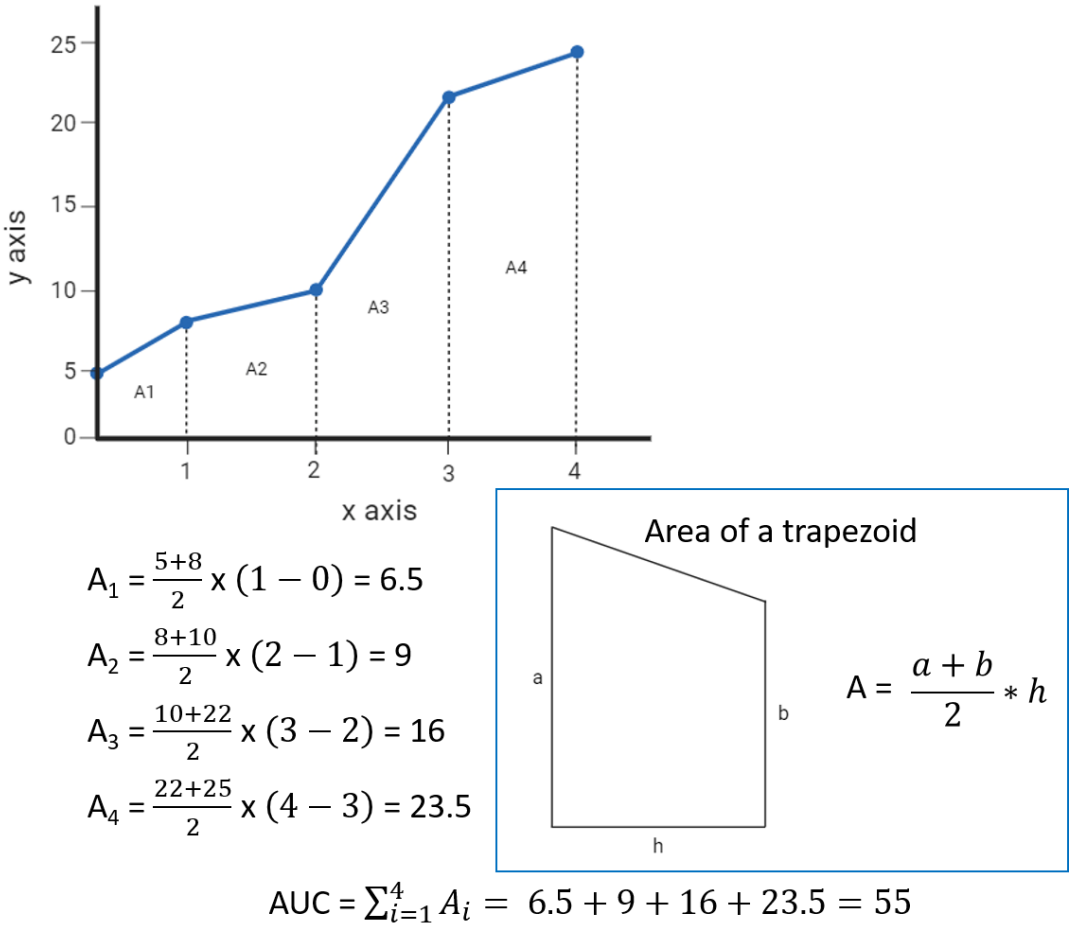


Figure 41. The process of calculating the area under the curve using the trapezoid rule. AUC: area under the curve.

4.4 Results

There is evidence in the literature that it is possible to safely reduce treatment intensity for patients with good risk genetics (Østergaard *et al.*, 2024). Findings from Chapter 3 also support this, as patients receiving 1 delayed intensification had optimal outcomes in both genetic subgroups. Due to the optimal outcomes seen on UKALL clinical trials for these patients, one can assume there exists an optimal treatment pathway within them that minimises intensity but maintains high cure rates. Studies have shown that drug doses can be used to determine intensity of treatment. Thus, in order to compare intensity of these

pathways, drug dosages were calculated and dose intensity scores developed to identify the optimal pathway for cure of *ETV6*::*RUNX1* and high hyperdiploidy patients.

4.4.1 Daily drug dosages for UKALLXI92, UKALL97, UKALL97/99, UKALL2003 and UKALL2011

A dataframe was created of the daily doses of the drugs given for each treatment pathway on the trials spanning the full treatment period. A heatmap for a selection of the treatment pathways is presented in Figure 42 which encompasses the possible variation seen on all the pathways. Data are presented for the 10 chemotherapeutic drugs administered on the protocols which are: steroid, vincristine, L-asparaginase, methotrexate, daunorubicin, etoposide, cytarabine, purine, cyclophosphamide, and doxorubicin. It is clear from the figure that there were major changes to treatment protocols between UKALL97 and UKALL97/99. UKALLXI92 and UKALL97 show similarities in dosage and timings of the administration overall, whilst the timings for each regimen on UKALL97/99, UKALL2003, and UKALL2011 are also largely similar to one another. This could partly explain the difference in outcome seen between these two groups in chapter 3. It is clear that these data are complex and not readily interpretable with regard to treatment intensity in this format. A method that encompasses all of these data into a single metric is required for any sort of meaningful analyses, hence the development of a dose intensity score was sought.

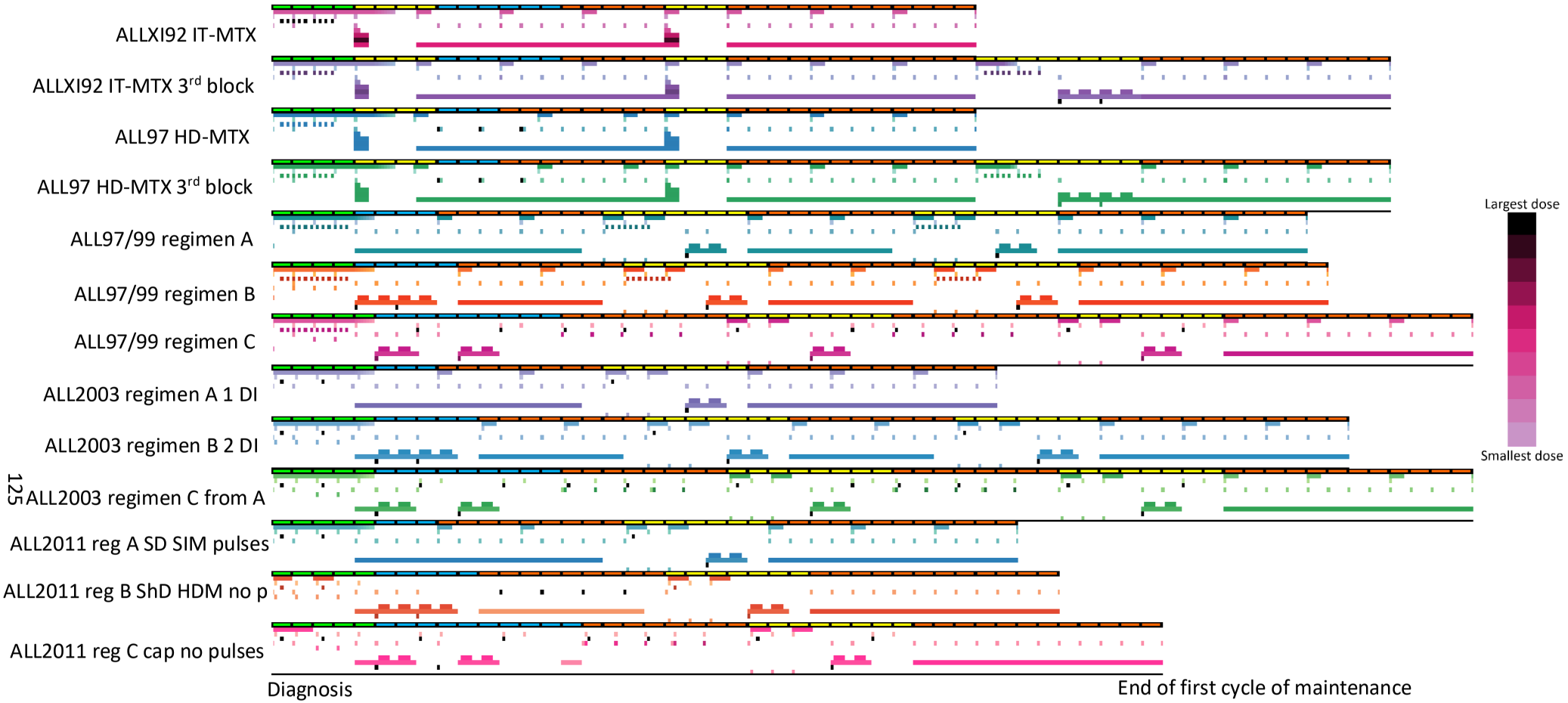


Figure 42. Heatmap of the daily drug dosages for patients on treatment pathways from UKALLXI92, UKALL97, UKALL97/99, UKALL2003, and UKALL2011. Magnitude of dosage is represented by colour intensity. The highest possible dose of escalating methotrexate during Capizzi interim maintenance was assumed and no escalation of purine was assumed. Intrathecal methotrexate doses for the two year age group are shown. Drugs are given for each trial by row in the following order: steroid, vincristine, L-asparaginase, methotrexate, daunorubicin, etoposide, cytarabine, purine, cyclophosphamide, and doxorubicin. Note that if the row is blank then that drug was not given on that treatment pathway. Maintenance was split into 12 week phases one of which is shown for each pathway. Total therapy ran for 100 weeks on UKALLXI92, 105 weeks on UKALL97, and 2 or 3 years from the start of interim maintenance 1 for girls or boys respectively on UKALL97/99, UKALL2003, and UKALL2011. Weeks are shown at the top each pathway and are indicated green for induction, blue for CNS therapy, yellow for intensifications and orange for maintenance/ interim maintenance. SD: standard dexamethasone, ShD: short dexamethasone, p: pulses, DI: delayed intensification, IT-MTX: intrathecal methotrexate, HDM-MTX: high dose methotrexate, reg: regimen.

4.4.2 Dose intensity

The dose intensity scores for 5,880 patients were calculated using the method described in Section 4.3.2. Patients were grouped by pathway to compare intensity score, where the mean of the score was used due to variations in dosages for patients resulting in multiple scores for one pathway [Supplementary Table 1]. These variations included different intrathecal methotrexate doses by age and escalating dosages of methotrexate and purine. There is currently no official method to compare the intensity of treatment pathways on different clinical trials. However, it is possible to compare the calculated dose intensity score with prior knowledge of intensities of pathways within an individual protocol. For example, it is apparent that pathways randomised to receive an additional intensification block would be more intensive than their counterparts. Thus, patients on UKALL2003 randomised to receive 1 DI should have a lower DIS than those randomised to receive 2 DIs. Furthermore, it is clear that the regimens on the latter trials have an ordering of intensity of A, B, and C from least to most intensive by design. Additionally, pathways that received high dose methotrexate for their consolidation therapy were more intensive than those that received intrathecal methotrexate. Finally, as girls received one year less of therapy than boys, it is to be assumed that the pathways for girls should always be assigned a lower score than the equivalent pathways for boys.

4.4.2.1 Dose intensity score

Using the above criteria, it was clear that the dose intensity score initially calculated was unsuccessful at determining the intensity of pathways. Whilst many of the UKALL2011 and UKALL2003 pathways were low on the scale and the high dose methotrexate pathways were near the top, regimen C was reported to be less intensive than the groups that received 2 DIs on regimens A and B within UKALL2003. Furthermore, within UKALL97/99, regimen B was determined to be more intensive than regimen C [Figure 43, Supplementary Table 1].

The hypothesis to explain the lack of success of this score was that there was such large variation in the baseline doses for each drug that a single additional administration of a drug with a large baseline dose (e.g. 200mg/m^2) could impact the score much more significantly than another with a lower baseline dose (e.g. 1.5mg/m^2) and thus skew the overall dose intensity score. A proposed solution to this issue was to calculate a dose intensity score in

which the relative doses of each drug (where the minimum dose of each drug is used as the baseline) are used in the formula. The results of this are given in Section 4.4.2.2.

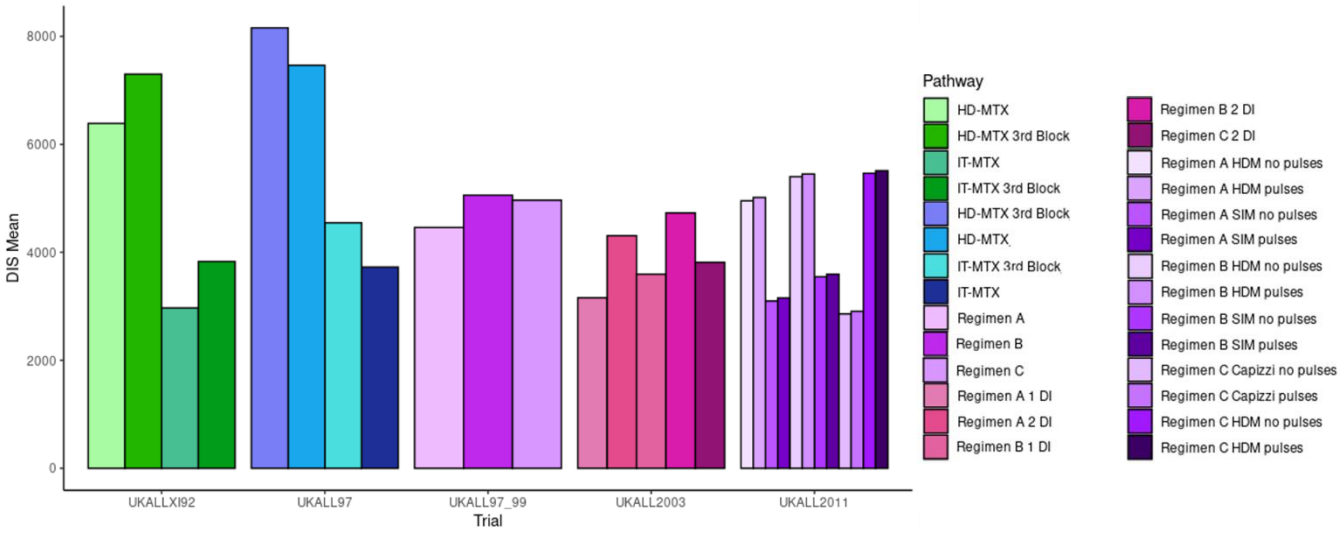


Figure 43. Bar chart of the dose intensity score for each pathway clustered by trial.

4.4.2.2 Relative dose intensity score

The same criteria as described in Section 4.4.2 were used to determine the accuracy of the relative dose intensity score. The range in mean values assigned to pathways by the relative score was 5.895517 – 111.104, which is a difference of 105.20848 in intensity; largely being driven by the stark difference in scores determined for the high dose methotrexate pathways and those receiving intrathecal methotrexate [Figure 44, Supplementary Table 1]. As expected, every pathway that received one delayed intensification had a lower dose intensity score than those receiving two delayed intensifications. Additionally, all of the high dose methotrexate pathways (administered in trials UKALLXI92, UKALL97, and UKALL2011) had the highest values as they received methotrexate doses in g/m^2, a great deal larger than the doses of other drugs and regular methotrexate doses that were typically given in mg/m^2. Furthermore, for each trial the regimens were ordered as regimen A, B and C from least to most intensive, and the pathways for additional intensifications were determined to be more intensive than their counterparts with fewer intensification blocks. For certain treatments, the pathway for girls was determined to be more intensive than the pathway for boys by mean score, however, this difference was always marginal (> 0.061) and likely due to differences in the age of patients affecting intrathecal methotrexate dose or random

assignment of escalating purine and methotrexate doses [Supplementary Table 1]. Therefore, initial analysis suggests that the hypothesis that relative doses were required in the model was true and the relative dose intensity score largely accurately determines intensity of treatment from drug dosages alone.

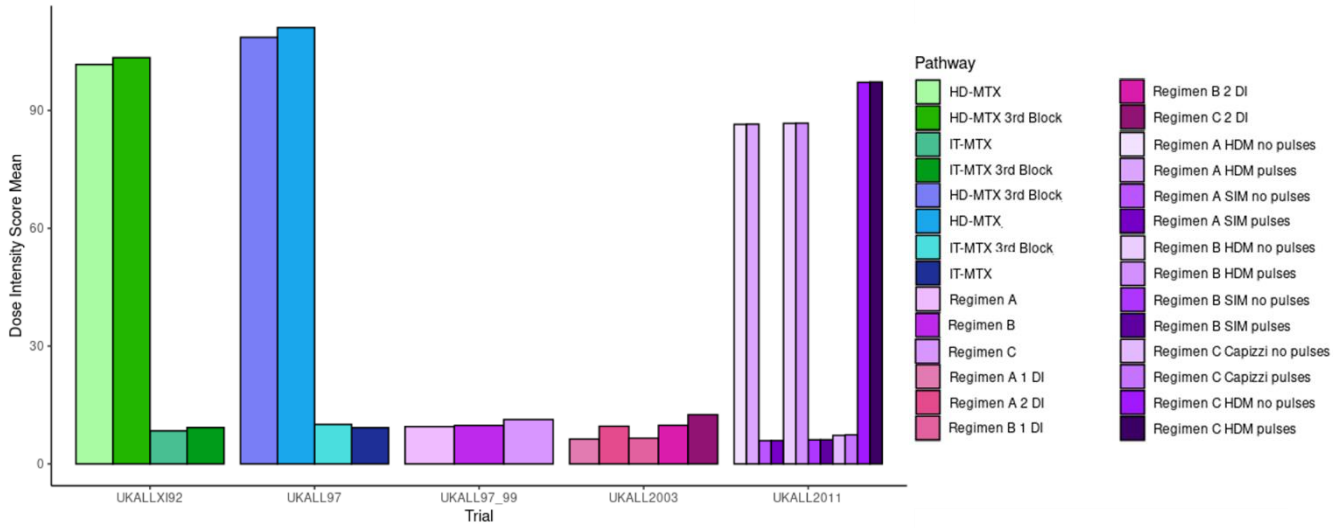


Figure 44. Bar chart of the relative dose intensity score for each pathway clustered by trial.

To determine if there was a difference in outcome by intensity of treatment, the relative dose intensity score was split into quartiles based on the full scale of values for all 5,880 patients for which it was calculated. This resulted in the groups: <8.48, 8.48 - 9.7506, 9.75061- 11.442, and ≥11.442. These four groups were deemed very low, low, intermediate, and high intensity respectively. This was done to compare the treatment arms that were considered very low intensity to the higher intensities directly rather than by a unit increase of intensity which does not provide information regarding optimal treatment elements. Furthermore, 5-year survival rates of the groups can be compared with this method to easily determine if cure rates were affected by treatment intensity. The survival of both *ETV6*::*RUNX1* and high hyperdiploidy patients were assessed by these groups.

Within *ETV6*::*RUNX1*, the very low risk group had the best overall survival with 5-year survival rates of 99% (95% CI (97-99)) compared to 96% (95% CI (93-98)), 97% (95% CI (95-98)), and 93% (95% CI (89-95)) for groups low, intermediate and high respectively. The difference in survival between the very low risk group and the low and high groups were significant with hazard ratios of 2.94, 95% CI (1.49-5.82), p = 0.002 and 3.59, 95% CI (1.79-7.23), p < 0.001

respectively. There was no difference in survival between the very low and intermediate group, nor the low and high groups (p = 0.179 and p = 0.481). The survival of the groups was not ordered by intensity as the intermediate group had better overall survival than the low group [Figure 45].

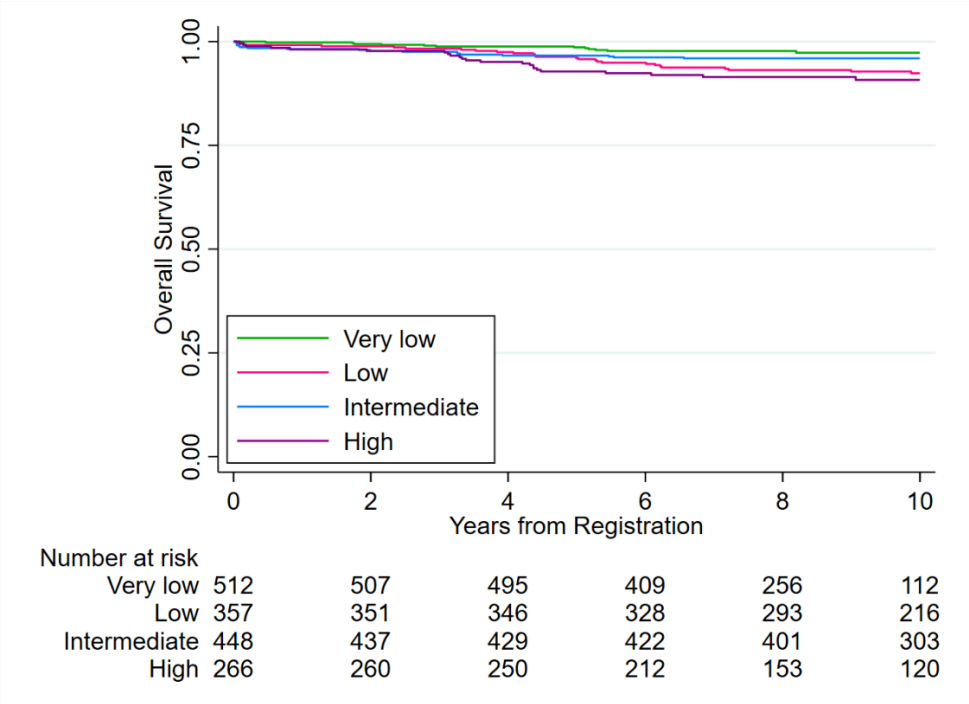


Figure 45. Overall survival of *ETV6*::*RUNX1* patients by relative dose intensity score quartile groups.

In terms of event-free survival, the outlook is very similar, with the very low intensity group once again having the best rates, followed closely by the intermediate group and then the low and high intensity groups. As shown in Figure 46, there is a more pronounced difference between the low and high groups demonstrated by the respective 5-year survival rates of 87% (95% CI (83-90)) and 81% (95% CI (76-85)). There was a significant difference in outcome for groups low and high when compared to the very low intensity group (HR: 1.84, 95% CI (1.24-2.74), p = 0.003 and log-rank p-value < 0.001), whereas the intermediate group had no increase in risk (HR: 1.13, 95% CI (0.75-1.72), p = 0.560).

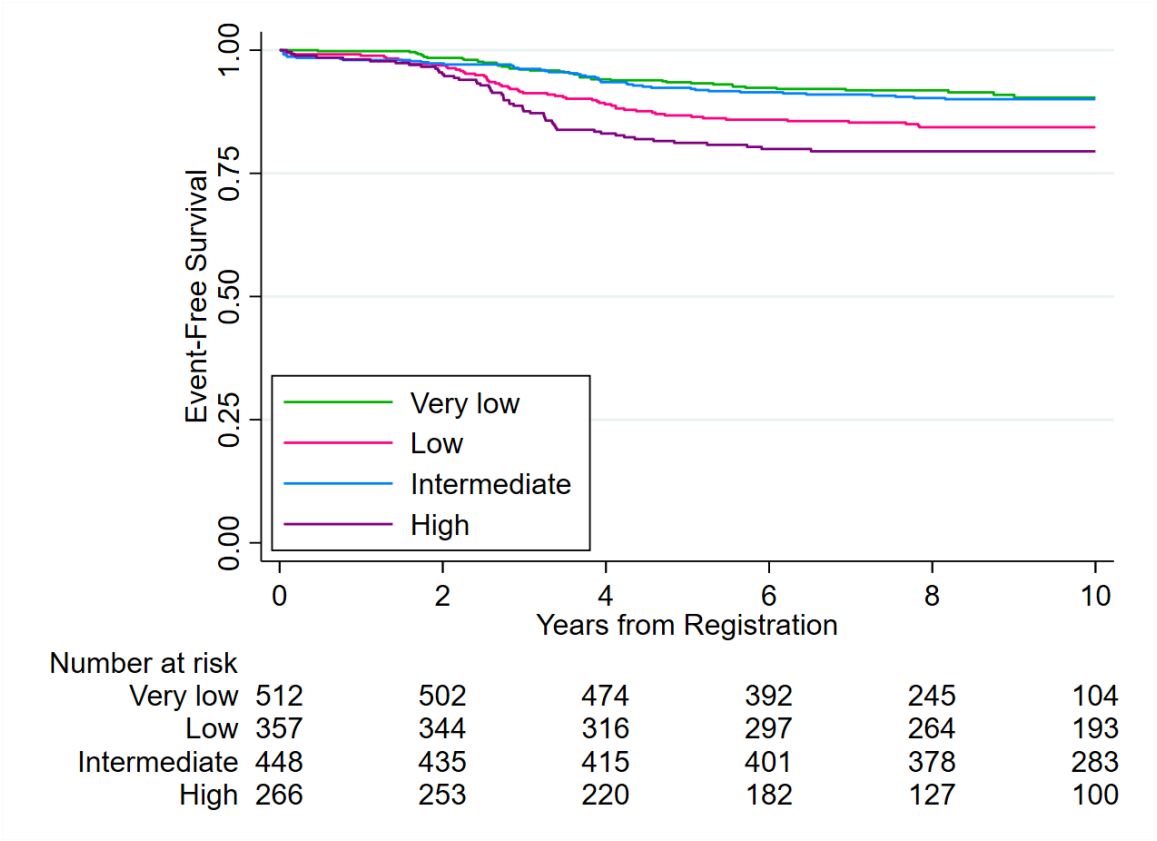


Figure 46. Event-free survival of *ETV6*::*RUNX1* patients by relative dose intensity score split into quartiles.

Survival analysis was performed within each trial as well as by NCI risk group to determine if the difference in cure rates by dose intensity quartiles were still evident within these subgroups. Statistics are only reported for groups with 10 or more patients to ensure strong statistical power and reliable findings. Log-rank p-values are presented in place of a hazard ratio if the Cox proportional hazards assumption was violated. Distribution of cases by trial and NCI risk across the relative dose intensity score quartiles were also assessed.

Distribution of cases by trial in the *ETV6*::*RUNX1* subgroup are presented in Table 35. As expected, the very low intensity group is mostly comprised of UKALL2003 and UKALL2011 patients (50% and 48% respectively), however there are also a large proportion of UKALL2003 patients within the low and intermediate intensity groups, whilst UKALL2011 patients are almost exclusively split between the very low and high intensity groups. The UKALLXI92, UKALL97, and UKALL97/99 patients are largely present in the low, intermediate, and high

intensity groups with the UKALL97 and UKALL97/99 patients predominantly being assigned as intermediate risk.

	Very low	Low	Intermediate	High
Total, n(%)	528 (33)	372 (23)	433 (27)	250 (16)
Trial				
UKALLXI92	9 (2)	50 (13)	0 (0)	64 (26)
UKALL97	0 (0)	38 (10)	102 (24)	20 (8)
UKALL97/99	2 (0.38)	94 (25)	91 (21)	5 (2)
UKALL2003	266 (50)	175 (47)	238 (54)	59 (24)
UKALL2011	251 (48)	15 (4)	2 (0.46)	102 (41)

Table 35. Distribution of *ETV6*::*RUNX1* cases by trial for the four relative dose intensity score groups. Pearson's $\chi^2 < 0.001$

Table 36 shows the distribution of cases by NCI risk in the *ETV6*::*RUNX1* subgroup. Interestingly, the proportion of patients by NCI group is identical in the low and intermediate group, whilst there are proportionally fewer NCI good risk patients in the very low and high intensity groups. It is clear that the majority of NCI good risk patients were assigned to the very low group and the fewest to the high intensity group, however the inverse is not seen in the NCI poor risk group which also had the majority of patients assigned to the very low risk. This suggests that any differences seen in outcome due to the relative dose intensity score groups is not being driven by NCI risk.

	Very low	Low	Intermediate	High
Total	528 (33)	372 (23)	433 (27)	250 (16)
NCI Risk				
NCI Good risk	404 (77)	309 (83)	359 (83)	162 (65)
NCI Poor risk	124 (23)	63 (17)	74 (17)	88 (35)

Table 36. Distribution of *ETV6*::*RUNX1* cases by NCI risk for the four relative dose intensity score groups. Pearson's $\chi^2 < 0.001$

There was no statistically significant difference in cure rates across the 4 quartile intensity groups in any of the trials for *ETV6*::*RUNX1* patients [Table 37]. Whilst not significant, cure rates increased with dose intensity within UKALLXI92, whilst the opposite was generally seen within the other trials, with the very low intensity group having the highest cure rates, as was seen in the overall analysis.

	Very low	Low	Intermediate	High
5-year overall survival rates (95% CI)				
UKALLXI92	-	92% (80-97)	-	94% (84-98)
UKALL97	-	100%	96% (90-99)	90% (66-97)
UKALL97/99	-	98% (92-99)	97% (90-99)	-
UKALL2003	99% (97-99.8)	96% (92-98)	96% (92-98)	95% (85-98)
UKALL2011	98% (96-99)	93% (61-99)	-	96% (90-99)
Hazard ratio (95% CI), p				
UKALLXI92	-	1.34 (0.50-3.56), p = 0.562	-	1
UKALL97	-	-	0.71 (0.15-3.45), p = 0.671	1
UKALL97/99	-	1.94 (0.48-7.74), p = 0.350	1	-
UKALL2003	0.44 (0.11-1.76), p = 0.247	1.26 (0.35-4.50), p = 0.726	0.84 (0.23-3.04), p = 0.787	1
UKALL2011	0.52 (0.14-1.92), p = 0.323	1.63 (0.18-14.63), p = 0.660	-	1

Table 37. 5-year overall survival rates and hazard ratios of *ETV6*::*RUNX1* patients by relative dose intensity score within the trials.

Within NCI risk the same trend as trial can be seen, with 5-year OS rates generally increasing as the dose intensity decreases. However, this difference is once again not significant in the NCI good risk group and largely insignificant in the NCI poor risk group [Table 38]. There was

a significantly reduced hazard in the very low intensity group compared to the high intensity group in the NCI poor risk subgroup however (HR: 0.28, 95% CI (0.09-0.90), p = 0.033).

	Very low	Low	Intermediate	High
5-year overall survival rates (95% CI)				
NCI Good risk	98% (96-99)	97% (94-98)	97% (95-98)	96% (92-98)
NCI Poor Risk	98% (94-99.6)	94% (84-98)	91% (81-95)	90% (81-95)
Hazard ratio (95% CI), p				
NCI Good risk	0.51 (0.21-1.24), p = 0.21	1.07 (0.49-2.35), p = 0.870	0.60 (0.25-1.39), p = 0.233	1
NCI Poor Risk	0.28 (0.09-0.90), p = 0.033	0.96 (0.37-2.52), p = 0.935	0.82 (0.31-2.17), p = 0.696	1

Table 38. 5-year overall survival rates and hazard ratios of *ETV6*::*RUNX1* patients by relative dose intensity score within the NCI risk groups.

As with overall survival, there is no difference in outcome between the intensity groups within any of the trials when looking at event-free survival. Whilst there was a general improvement in 5-year EFS rates as the intensity decreased, it was not as pronounced as in OS, with UKALL97/99 patients having the best rates in the intermediate intensity group and the UKALL2011 patients having superior rates in the low intensity group. However, these differences were minimal and non-significant as evidenced by the hazard ratios in Table 39.

	Very low	Low	Intermediate	High
5-year event-free survival rates (95% CI)				
UKALLXI92	-	64% (49-76)	-	64% (51-74)
UKALL97	-	89% (74-96)	86% (78-92)	80% (78-92)
UKALL97/99	-	91% (84-96)	93% (86-97)	-
UKALL2003	95% (92-97)	91% (86-95)	94% (90-96)	90% (79-95)
UKALL2011	92% (88-95)	93% (61-99)	-	90% (82-95)
Hazard ratio (95% CI), p				
UKALLXI92	-	0.87 (0.48-1.59), p = 0.658	-	1
UKALL97	-	p = 0.948	0.85 (0.29-2.51), p = 0.766	1
UKALL97/99	-	1.09 (0.42-2.83), p = 0.856	1	-
UKALL2003	0.57 (0.24-1.36), p = 0.201	0.88 (0.37-2.11), p = 0.776	0.67 (0.28-1.59), p = 0.363	1
UKALL2011	0.94 (0.45-1.98), p = 0.877	0.63 (0.08-4.96), p = 0.665	-	1

Table 39. 5-year event-free survival rates and hazard ratios of *ETV6*::*RUNX1* patients by relative dose intensity score within the trials.

For both NCI good and poor risk patients, the very low and intermediate intensity groups have significantly better event-free survival than the high intensity group, whilst the low intensity group had comparable rates as was seen in the overall analysis [Table 40]. Whilst not significant, there was a more pronounced difference in 5-year EFS rates between the high and low intensity groups in the NCI poor risk patients than the NCI good risk patients (74% vs 86% and 86% vs 88% respectively).

	Very low	Low	Intermediate	High
5-year event-free survival rates (95% CI)				
NCI Good risk	93% (90-95)	88% (84-91)	93% (89-95)	86% (79-90)
NCI Poor Risk	92% (85-96)	86% (74-92)	89% (80-94)	74% (63-82)
Hazard ratio (95% CI), p				
NCI Good risk	0.51 (0.31-0.85), p = 0.010	0.87 (0.54-1.41), p = 0.578	0.55 (0.33-0.91), p = 0.021	1
NCI Poor Risk	0.38 (0.20-0.74), p = 0.004	0.48 (0.22-1.04), p = 0.062	0.46 (0.22-0.96), p = 0.040	1

Table 40. 5-year event-free survival rates and hazard ratios of *ETV6*::*RUNX1* patients by relative dose intensity score within the NCI risk groups.

The proportion of specific outcomes were compared across the intensity groups within *ETV6*::*RUNX1* to determine which type of event was prevalent in each group [Table 41]. The aim of this was to test the hypothesis that there would be fewer remission deaths in the lower intensity groups due to less toxicity and fewer relapses in the higher intensity groups due to more treatment. However, whilst there was a difference in the distribution ($p < 0.001$ by Pearson χ^2 test), there is no evidence of a association between lesser intensity and fewer remission deaths, nor increased intensity and fewer relapses.

	Total	Very low	Low	Intermediate	High
Outcomes					
Continuing remission	1387 (88)	470 (92)	303 (85)	403 (90)	211 (79)
Died in remission	19 (1)	5 (1)	3 (1)	7 (2)	4 (2)
Relapse/ refractory 2nd rem	115 (7)	30 (6)	27 (8)	26 (6)	32 (12)
Relapse/ refractory death	62 (4)	7 (1)	24 (7)	12 (3)	19 (7)

Table 41. The frequency and proportion of outcomes by relative dose intensity score groups for *ETV6*::*RUNX1* patients. Pearson $\chi^2 < 0.001$.

Within the high hyperdiploidy subgroup, the very low intensity group had significantly better overall survival when compared to the low, intermediate and high intensity groups with log-rank p-values: p = 0.0184, p = 0.0012, and p = 0.0329 respectively. The low intensity group had a 5-year survival rate of 98% (95% CI (96-99)) compared to ~94% for the other groups

[Figure 47]. There was no significant difference in survival between the other three groups with p > 0.2 by log-rank test for all comparisons. Again, there wasn't the expected correlation between survival and intensity as the low and intermediate groups had practically identical survival rates.

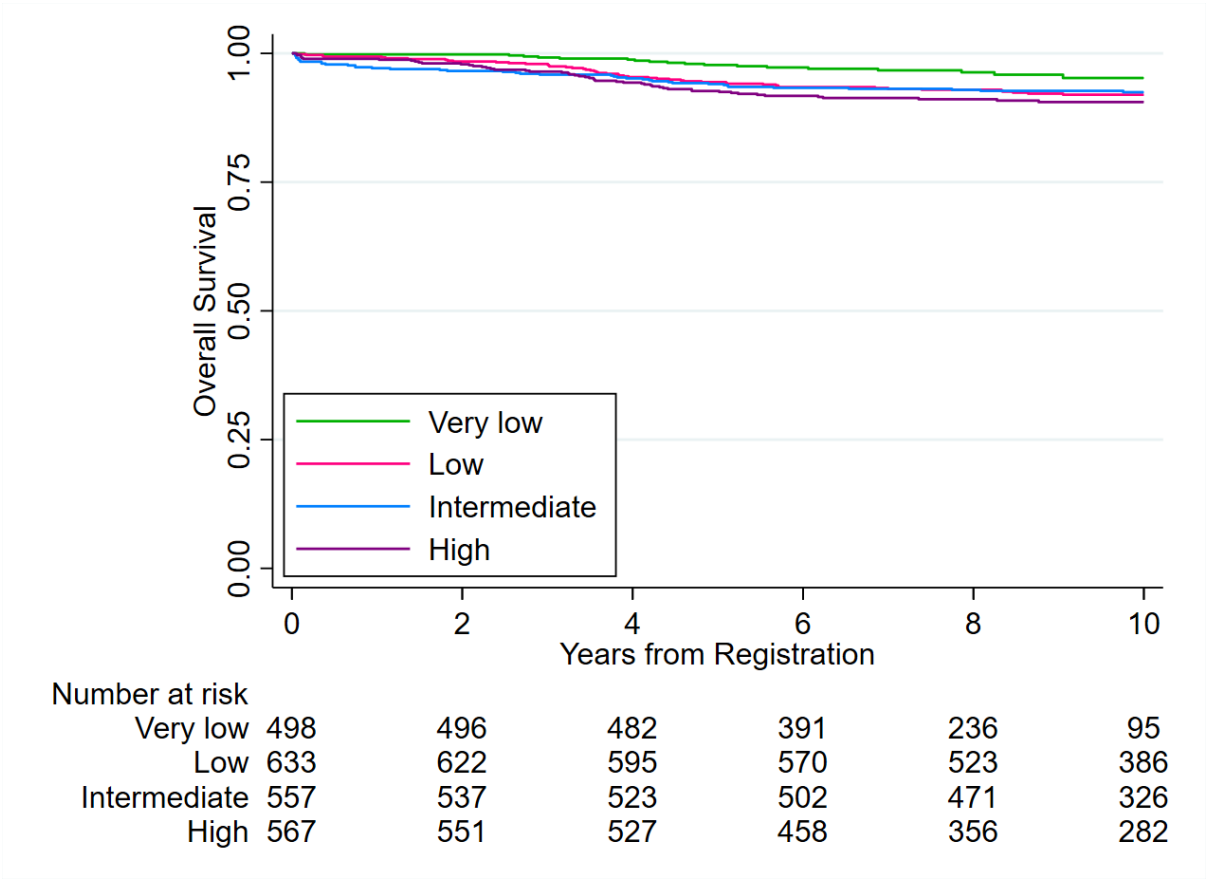


Figure 47. Overall survival of high hyperdiploidy patients by relative dose intensity score split into quartiles.

5-year survival rates were also superior for the very low intensity group in the event-free setting for high hyperdiploidy patients, at 92% (95% CI (90-94)) compared to 84%, 87%, and 81% in the low, intermediate and high groups respectively [Figure 48]. This difference was significant with hazard ratios of 2.12 (p < 0.001), 1.65 (p = 0.007), and 2.43 (p < 0.001) when compared to each group in order. The high intensity group also had inferior survival when compared to the intermediate risk group (HR: 1.48, 95% CI (1.11-1.96), p = 0.007) but there was no significant difference in survival between the intermediate and low intensity groups

(p = 0.087) or the low and high intensity groups (p = 0.281). Once again, the intermediate intensity group had better outcomes than the low intensity risk group.

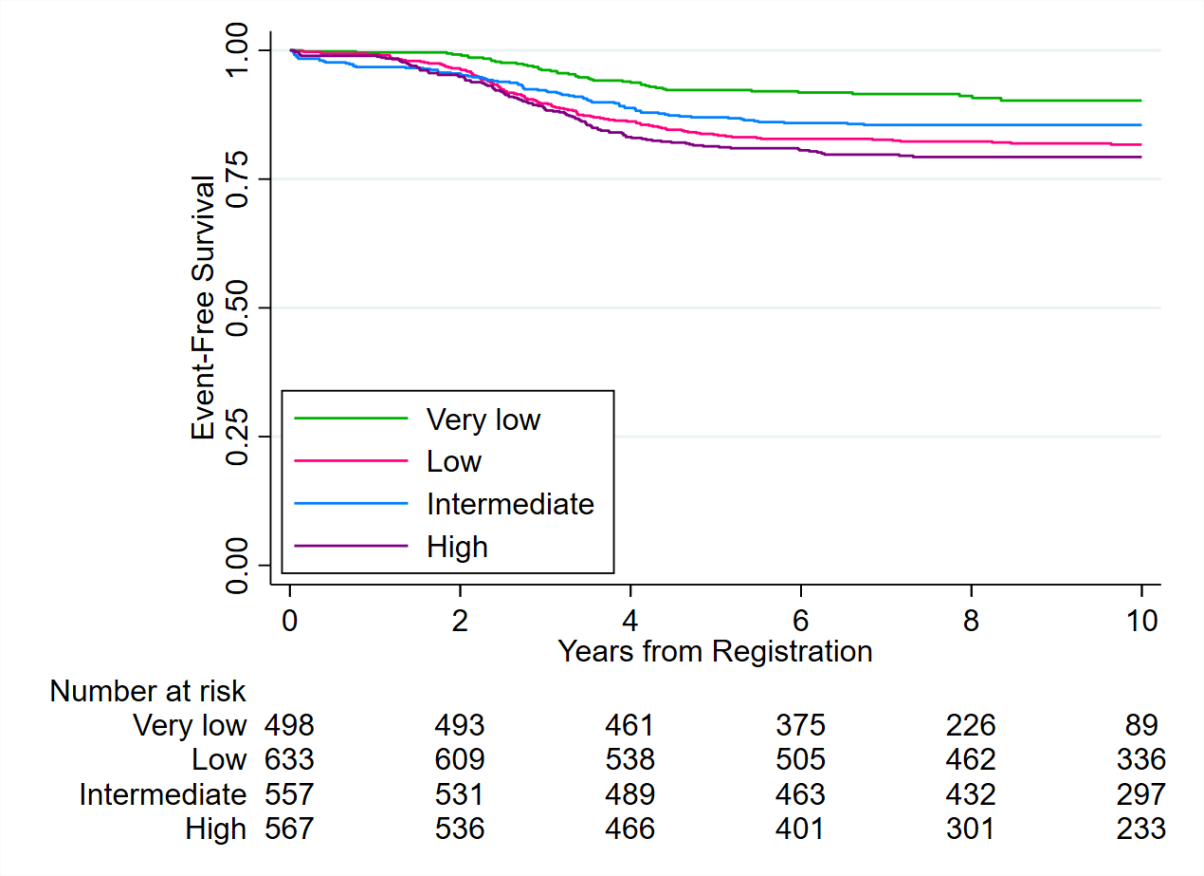


Figure 48. Event-free survival of high hyperdiploidy patients by relative dose intensity score split into quartiles.

Distribution of cases by trial in the high hyperdiploidy subgroup are presented in Table 42. Similarly to *ETV6*::*RUNX1* patients, the very low intensity group is mostly comprised of UKALL2003 and UKALL2011 patients, with a similar number of UKALL2003 patients assigned to the 3 lower intensity groups. The UKALL2011 patients are again primarily split between the very low and high intensity groups, which is due to the presence or absence of high dose methotrexate in the treatment arms. The same pattern seen in the *ETV6*::*RUNX1* subgroup for the UKALLXI92, UKALL97, and UKALL97/99 patients is evident in high hyperdiploidy patients, with the majority of UKALL97 and UKALL97/99 patients stratified to the low and intermediate intensity groups, and most UKALLXI92 patients assigned to the low and high intensity groups.

	Very low	Low	Intermediate	High
Total, n(%)	551 (24)	626 (28)	561 (25)	517 (23)
Trial				
UKALLXI92	43 (8)	141 (23)	0 (0)	189 (37)
UKALL97	2 (0.36)	91 (15)	144 (25)	25 (5)
UKALL97/99	6 (1)	134 (21)	120 (21)	4 (1)
UKALL2003	219 (40)	241 (39)	283 (50)	130 (25)
UKALL2011	281 (51)	19 (3)	14 (3)	169 (33)

Table 42. Distribution of high hyperdiploidy cases by trial for the four relative dose intensity score groups. Pearson's $\chi^2 < 0.001$

Distribution of cases by NCI risk in the high hyperdiploidy subgroup are presented in Table 43. Unlike in the *ETV6*::*RUNX1* patients, the proportion of NCI poor risk cases generally increases as the intensity increases. This could suggest that the superior outcome seen in the very low dose intensity group could be due to the high proportion of NCI good risk patients; however, the low intensity group which has the largest amount of NCI good risk patients, both in terms of numbers and proportion, has the second most inferior outcomes contradicting this notion.

	Very low	Low	Intermediate	High
Total	551 (24)	626 (28)	561 (25)	517 (23)
NCI Risk				
NCI Good risk	458 (83)	545 (87)	438 (78)	362 (70)
NCI Poor risk	93 (17)	81 (13)	123 (22)	155 (30)

Table 43. Distribution of high hyperdiploidy cases by NCI risk for the four relative dose intensity score groups. Pearson's $\chi^2 < 0.001$

When comparing these groups within the trials, there was once again a mostly negative association between dose intensity and cure rates, with overall survival improving with decreasing intensity in UKALLXI92 and UKALL2003 as well as stable rates in UKALL97/99. Cure rates largely increased in UKALL97, however the intermediate intensity group had marginally better OS at 5-years than the low intensity group. In UKALL2011, the high intensity and very

low intensity groups had comparable rates, as did the low and intermediate intensity groups. However, none of these differences by dose intensity groups within the trials were significant for high hyperdiploidy patients [Table 44]. The exception to this was in UKALL2003, where the very low intensity group had a significantly improved outcome compared to the high risk group (log-rank p-value = 0.01).

	Very low	Low	Intermediate	High
5-year overall survival rates (95% CI)				
UKALLXI92	93% (80-98)	91% (85-95)	-	89% (84-93)
UKALL97	-	92% (85-96)	94% (88-97)	88% (67-96)
UKALL97/99	-	96% (90-98)	96% (90-98)	-
UKALL2003	99% (96-99.8)	96% (92-98)	94% (90-96)	92% (86-96)
UKALL2011	97% (94-99)	94% (63-99)	93% (59-99)	98% (94-99)
Hazard ratio (95% CI), p				
UKALLXI92	0.64 (0.23-1.84), p = 0.410	1.07 (0.61-1.88), p = 0.811	-	1
UKALL97	-	0.66 (0.21-2.10), p =0.478	0.50 (0.16-1.56), p = 0.235	1
UKALL97/99	-	0.73 (0.29-1.85), p = 0.509	1	-
UKALL2003	p = 0.01	0.58 (0.26-1.26), p = 0.168	0.81 (0.40-1.64), p = 0.553	1
UKALL2011	2.43 (0.68-8.60), p = 0.170	3.21 (0.33-30.90), p = 0.312	-	1

Table 44. 5-year overall survival rates and hazard ratios of high hyperdiploidy patients by relative dose intensity score within the trials.

There was no significant difference in overall survival by dose intensity group in either NCI good or poor risk subgroups. The NCI good risk patients had 5-year OS rates of ~95% in all intensity groups [Table 45]. In the NCI poor risk subgroup, patients had OS rates of ~88% in the high, intermediate, and low intensity groups and a superior survival of 97% in the very

low intensity group. The difference between the high and very low group wasn't significant however with a hazard ratio of 0.38, 95% CI (0.13-1.14), p = 0.086.

	Very low	Low	Intermediate	High
5-year overall survival rates (95% CI)				
NCI Good risk	98% (96-99)	95% (93-96)	96% (94-98)	94% (91-96)
NCI Poor Risk	97% (90-99)	89% (80-94)	87% (80-92)	89% (83-93)
Hazard ratio (95% CI), p				
NCI Good risk	Log-rank p-value = 0.123	0.90 (0.56-1.43), p = 0.654	0.65 (0.38-1.10), p = 0.110	1
NCI Poor Risk	0.38 (0.13-1.14), p = 0.086	1.18 (0.55-2.53), p = 0.662	1.42 (0.74-2.72), p = 0.298	1

Table 45. 5-year overall survival rates and hazard ratios of high hyperdiploidy patients by relative dose intensity score within the NCI risk groups. The log-rank p-value is given in instances where the proportional hazards assumption was violated.

When comparing the event-free survival of the intensity groups within the trials, there were no significant differences, with the exception of UKALL2003. 5-year EFS rates improved as the intensity decreased, and the outcomes were significantly better in the low and very low intensity groups with p-values of 0.046 and 0.001 respectively [Table 46]. In UKALLXI92, the very low intensity group had inferior 5-year EFS rates than the high intensity group, however this difference wasn't significant with a hazard ratio of 1.24, 95% CI (0.73-2.12), p = 0.426.

	Very low	Low	Intermediate	High
5-year event-free survival rates (95% CI)				
UKALLXI92	65% (49-77)	70% (62-77)	-	70% (63-76)
UKALL97	-	84% (74-90)	86% (79-91)	80% (58-91)
UKALL97/99	-	87% (80-92)	83% (75-89)	-
UKALL2003	96% (93-98)	92% (88-95)	89% (84-92)	86% (79-91)
UKALL2011	90% (86-93)	78% (51-91)	85% (52-96)	90% (85-94)
Hazard ratio (95% CI), p				
UKALLXI92	1.24 (0.73-2.12), p = 0.426	0.94 (0.64-1.37), p = 0.732	-	1
UKALL97	-	-	0.81 (0.31-2.11), p = 0.660	1
UKALL97/99	-	0.72 (0.39-1.34), p = 0.299	1	-
UKALL2003	Log-rank p-value = 0.001	0.54 (0.30-0.99), p = 0.046	0.73 (0.43-1.27), p = 0.267	1
UKALL2011	0.52 (0.14-1.92), p = 0.323	1.63 (0.18-14.63), p = 0.660	-	1

Table 46. 5-year event-free survival rates and hazard ratios of high hyperdiploidy patients by relative dose intensity score within the trials. The log-rank p-value is given in instances where the proportional hazards assumption was violated.

Within both NCI risk groups, there was no significant difference in the event-free survival of the high and low intensity groups for high hyperdiploidy patients [Table 47]. In the NCI good risk subgroup, the very low and intermediate groups had significantly better EFS rates than the high risk group with hazard ratios of 0.49, 95% CI (0.34 – 0.72), $p < 0.002$ and 0.68, 95% CI (0.48 – 0.96), $p = 0.027$ respectively. This was also seen in the analysis of the intensity groups overall.

	Very low	Low	Intermediate	High
5-year event-free survival rates (95% CI)				
NCI Good risk	91% (88-94)	85% (82-88)	88% (85-91)	83% (79-87)
NCI Poor Risk	87% (78-92)	81% (71-88)	82% (74-88)	77% (69-83)
Hazard ratio (95% CI), p				
NCI Good risk	0.49 (0.34-0.72), p < 0.001	0.85 (0.62-1.16), p = 0.315	0.68 (0.48-0.96), p = 0.027	1
NCI Poor Risk	0.59 (0.32-1.08), p = 0.089	0.75 (0.42-1.35), p = 0.342	0.75 (0.44-1.26), p =0.274	1

Table 47. 5-year event-free survival rates and hazard ratios of high hyperdiploidy patients by relative dose intensity score within the NCI risk groups.

For high hyperdiploidy, the proportion of specific outcomes were again compared across the intensity groups to determine which type of event was prevalent in each group [Table 48]. There was a difference in the distribution (p = 0.002 by Pearson χ^2 test), however this difference was not driven by any specific association between intensity and outcome, as the low and high intensity groups had similar proportions of relapse (as did the very low and intermediate groups), and there was no evidence of a difference in the proportion of remission deaths.

	Total	Very low	Low	Intermediate	High
Outcomes					
Continuing remission	1905 (84)	455 (91)	516 (82)	479 (86)	455 (80)
Died in remission	30 (1)	2 (0.4)	9 (1)	10 (2)	9 (2)
Relapse/ refractory 2nd rem	183 (8)	24 (5)	64 (10)	36 (6)	59 (10)
Relapse/ refractory death	137 (6)	17 (3)	44 (7)	32 (6)	44 (8)

Table 48. The frequency and proportion of outcomes by relative dose intensity score groups for high hyperdiploidy patients. Pearson χ^2 = 0.002.

Whilst the relative dose intensity score met the criteria proposed to determine its functionality, there is no way to ascertain the true accuracy of the scores assigned by this model. As such, an alternative method of calculating the intensity of treatment using an area

under the curve method, proposed by Allgoewer *et al.*, was explored to compare the findings of both metrics.

4.4.2.3. Area under the curve dose intensity score

The functionality of the area under the curve dose intensity score was assessed. According to this score, the addition of pulses during maintenance in UKALL2011 caused a larger difference in intensity compared to the patients who didn't receive pulses, than the difference in intensity seen between regimens A and C on this same trial [Figure 49]. It is unclear if the intensity of low level maintenance therapy for a year is more than that of the higher intensity treatment elements such as high dose methotrexate that are given for a much shorter period and thus, it isn't possible to determine the functionality of the area under the curve dose intensity score method from this information alone. However, this score suggests that for patients on regimen A of UKALL2003, receiving one delayed intensification was more intensive than receiving two delayed intensifications – a finding which is known to be false. For this score, almost every pathway for the girls is ranked below every pathway for the boys [Supplementary Table 1]. This is because the curves are longer for the boys due to an additional year of maintenance therapy, which has a great effect on the score when this method of calculation is used. Therefore, it appears that the area under the curve dose intensity score was not able to successfully determine the intensity of treatment for ALL patients based on drug dosages alone.

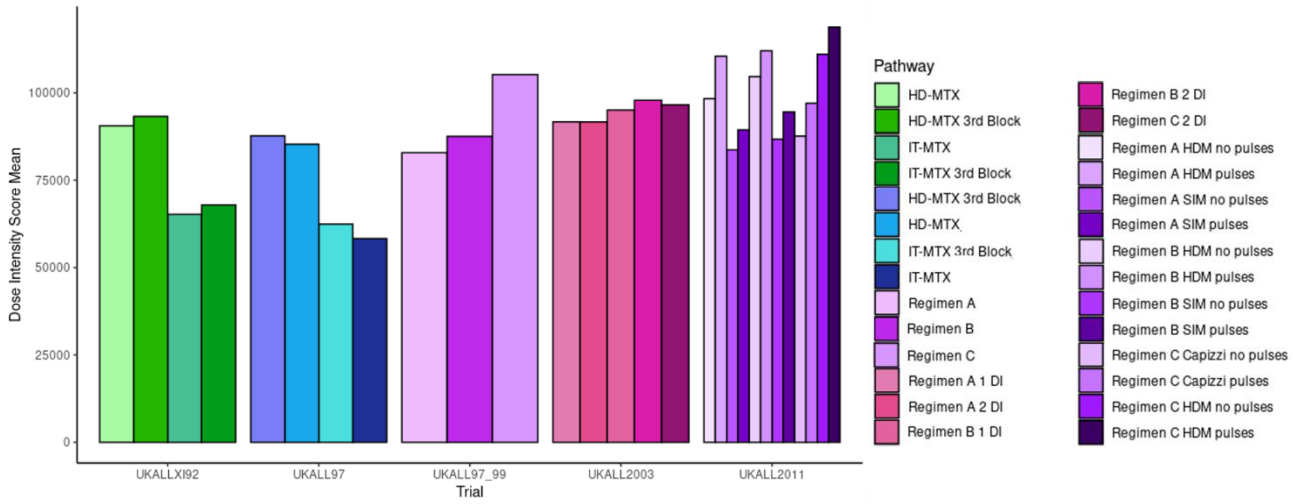


Figure 49. Bar chart of the area under the curve dose intensity score for each pathway clustered by trial.

4.5 Discussion

In summary, this chapter presents the drug dosage data for four paediatric UK clinical trials – data that were previously unavailable, and assesses the functionality of three different methods for calculating a dose intensity score. The analysis performed in this chapter identified an effective formula for calculating the intensity of treatment within paediatric ALL. It also determined that intensity of treatment was largely prognostic within these patients, with a negative correlation between intensity and outcome.

Within the context of these data, the dose intensity score proposed by Hryniuk *et al.* and the area under the curve method adapted from Allgoewer *et al.* were unsuccessful at determining intensity of treatment pathways. The dose intensity score likely failed as the differences in standard dosages for each drug varied immensely from 1.5mg/m^2 in vincristine to 1000mg/m^2 for cyclophosphamide, before factoring in the 6 or 8g/m^2 of methotrexate patients assigned to the high dose methotrexate arms received. Thus, if the standard doses of the drugs given during maintenance therapy were higher than those received during delayed intensifications for example, then this interruption would actually result in a lower intensity score overall – something known to be false. The area under the curve method can be affected by this same issue, as the total daily doses for the period of the delayed intensification would be smaller resulting in a lower curve and thus a smaller area – a matter which is avoided when using relative doses. Moreover, the additional year of therapy for boys on the latter trials affects the output of the area under the curve dose intensity score as the curve is longer overall resulting in a much larger area, which can surpass the area increase caused by known high intensity treatment elements such as delayed intensifications and HD-MTX. This issue is avoided in the method used to calculate the dose intensity score and relative dose intensity score as the values are divided by the number of weeks per phase, thus creating a weekly average score and lessening the effect of the increased maintenance length.

Receiving high dose methotrexate during therapy was the most intensive form of treatment regardless of number of delayed intensifications or regimen, with patients receiving one delayed intensification and HDM interim maintenance on UKALL2011 ranking as more intensive than receiving two or three delayed intensifications and standard intrathecal methotrexate on any other trial. Furthermore, the very low intensity group had the best outcome in both overall and event-free survival for *ETV6*::*RUNX1* and high hyperdiploidy

patients suggesting that the less intensive therapy is sufficient in these subgroups. Moreover, the HD-MTX treatment arms were all classified as high intensity – the group with the lowest survival rates. This suggests that high dose methotrexate is not necessary, and could potentially be negatively impacting survival in the good risk genetics subgroups, which corroborates the findings of Østergaard *et al.* who determined that HD-MTX and pulses of glucocorticoids and vincristine in maintenance were superfluous (Østergaard *et al.*, 2024). The results also suggest that receiving only one intensification is sufficient in treatment of good risk genetic ALL subgroups. Therefore, future protocols should treat these patients with standard interim maintenance and one delayed intensification regardless of NCI risk.

There was no correlation between relapse and treatment intensity, highlighting that low intensity treatment did not cause an increase in relapses within the good risk genetic subgroups [Tables 39 and 44]. This suggests that de-escalation is feasible within these populations. There was also no association found when analysing remission deaths and treatment intensity suggesting that no treatment pathway in this cohort was intensive enough to cause toxic death. This does not exclude the possibility of toxicities and long-term late effects caused by this treatment however, and thus de-escalated treatment should be standard for good risk patients.

Outcomes in all end points for both subgroups seem to suggest that intensity of treatment according to the relative dose intensity score is prognostic, with the very low intensity group displaying the highest cure rates and event-free survival rates. This is further supported by the fact that, whilst not significant, there was a largely negative correlation between treatment intensity and survival within trials and NCI risk groups; thus demonstrating a prognostic effect independent from these factors.

Although the relative dose intensity score appears to successfully determine intensity of patients across multiple trials, there is no method to determine the true accuracy of this score. Furthermore, whilst the doses of HDM-MTX given are much higher than the doses of other drugs administered, the effect this has on the relative dose intensity score (> 10 times score than their IT-MTX counterparts) is unlikely to reflect the actual intensity of these treatment pathways. Thus, it is possible that a more sophisticated method of assessing the effect of dose could accurately predict the outcome of patients. Therefore, the next stage of

this study was to apply machine learning methods to identify optimal dose for cure, which is the focus of chapter 5.

Chapter 5. Utilisation of machine learning methods to identify optimal treatment elements for ALL patients with good risk genetics

5.1 Introduction

Research efforts in cancer focus largely on etiology and therapy, where the large accumulation of cancer-associated data are providing insights into the causes of cancer and the mechanisms involved in its progression; as well as the ability to develop and adapt treatment based on risk factors and response (Elmore *et al.*, 2021; Sebastian and Peter, 2022). Until recently, traditional statistical methods have been the main approach to this research, with artificial intelligence largely restricted to computer vision tasks such as using imaging to diagnose malignancies (Elemento *et al.*, 2021). However, thanks to technological advances and an increased digitisation of patient data through electronic health records, artificial intelligence and machine learning methods are emerging in cancer research, with a vast number of applications which are continually increasing (Sarker, 2021; Cuocolo *et al.*, 2020). Some key areas of cancer research in which machine learning techniques have promising applications are in cancer detection, subtype classification, optimisation of treatment, and personalising therapy (Elemento *et al.*, 2021).

In terms of cancer detection and classification, deep neural networks have been utilised in multiple studies to distinguish cancer cells from healthy cells or determine specific cancer subtypes. For example, Ehteshami Bejnordi *et al.* were able to develop algorithms that accurately detect lymph node metastases in women with breast cancer with a comparable accuracy to pathologists (Ehteshami Bejnordi *et al.*, 2017). Deep convolutional neural networks were used to discriminate between two subtypes of lung cancer, four biomarkers of bladder cancer, and five biomarkers of breast cancer, as well as accurately grade hepatocellular carcinoma nuclei (Khosravi *et al.*, 2018; Li, Jiang and Pang, 2017). Another study was able to identify prostate cancer and distinguish it from other benign conditions (Wang *et al.*, 2017).

In several studies, machine learning methods have also been utilised with the aim of predicting patient survival. One study developed a machine learning based model to discriminate risk of patients with breast cancer based on mammograms and traditional risk factors (Yala *et al.*, 2019). Through the use of machine learning algorithms on MRIs, tumour surface regularity was found to be independently prognostic in glioblastoma, whilst radiomic features were associated with progression-free survival in nasopharyngeal carcinoma (Pérez-Beteta *et al.*, 2018; Zhang *et al.*, 2017). Furthermore, a deep learning method was shown to

outperform both the Cox proportional hazards model and the random survival forest algorithm in predicting survival of oral cancer patients, whilst a decision support system was able to predict the progression-free survival of breast cancer patients (Kim *et al.*, 2019; Ferroni *et al.*, 2019).

Machine learning also has applications in therapy response and drug efficacy. Machine learning algorithms were employed to identify patients sensitive to a specific drug combination in treating colorectal cancer, and patient response to intra-arterial therapies of hepatocellular carcinoma were successfully predicted using a supervised learning approach (Lu *et al.*, 2020; Abajian *et al.*, 2018). Further studies assessing drug sensitivity were reviewed by Liang *et al.* which included studies predicting the drug sensitivity of patients with ovarian cancer, gastric cancer, and endometrial cancer, as well as studies assessing drug resistance in cancer (Liang *et al.*, 2020).

Finally, an area of research in which machine learning methods are gaining importance is that of calculating drug dosages and optimal drug combinations. Several studies have been published using methods such as linear regression, support vector machines, random forest, and convolutional neural networks (Banegas-Luna *et al.*, 2021). However, the use of ML in this area remains limited and the use of these approaches in leukaemia still need to be addressed.

Machine learning methodologies are employed in this chapter to predict outcome of patients based on drug dosages and treatment elements as an alternative approach to the traditional statistical methods explored in the previous chapters. These traditional statistical approaches had several limitations, largely due to the fact that these approaches cannot effectively capture complex relationships within data. As such, it was difficult to assess the effect of different treatment elements with these methods, as there was no way to determine if the difference in outcomes was due to one specific treatment element alone or as a result of a combination of elements. Furthermore, whilst the relative dose intensity score appeared to accurately calculate treatment intensity for patients, it could not account for interactions between drugs affecting overall intensity. The benefit of using machine learning for these tasks is that the algorithms have the ability to find statistical patterns in large, complex data and use these to make predictions that could not be easily explained or understood through

traditional statistics. Examples of this include the ML models utilised in this project, namely the decision tree, random forest, and XGBoost algorithms which are easily interpretable, presenting a clear understanding of how predictions are made compared to complicated multivariate models in traditional statistics.

5.2 Aims

The aims of this chapter are to:

- Produce a machine learning algorithm which accurately predicts outcome of good risk genetics patients based on drug dosage.

- Identify optimal treatment elements for good risk genetics patients based on the drug dosage thresholds defined by the machine learning algorithm.

- Validate the findings of the relative dose intensity score.

5.3 Methods

Classification decision trees were utilised within both *ETV6*::*RUNX1* and high hyperdiploidy data separately, with three different target variables considered. These variables were considered as they are the outcomes of interest for risk stratification of ALL. The first target variable was a 4-class outcome variable: continuing remission, remission death, relapse or refractory disease (denoted rel/ref) leading to 2nd remission, and rel/ref leading to death. The second target variable was an indicator variable for remission death (i.e. remission death yes/ no) and similarly the final target variable considered was an indicator variable for rel/ref regardless of outcome. Decision trees with different features were considered which were named Chestnut, Oak, and Elm. Chestnut had the features: trial and the total drug dosages where the drugs were: cyclophosphamide, etoposide, vincristine, L-asparaginase, steroid, methotrexate, purine, anthracycline, and cytarabine. Oak had the features: trial, regimen, delayed intensifications, purine received, and steroid received. Thus, Oak was only employed on data for the trials that were stratified by regimen: UKALL97/99, UKALL2003, and UKALL2011. Elm had the same features as Oak minus regimen so that all the trials could be considered by these features (trial, delayed intensifications, purine received, and steroid

received). This information is summarised in Table 49. It was required for features to be numerical and thus, the string variable trial was converted with the following coding: 0 = UKALLXI92, 1 = UKALL97, 2 = UKALL97/99, 3 = UKALL2003, 4 = UKALL2011. Random forest and XGBoost algorithms were only utilised with the features from the Chestnut tree as these were shown to be the best at accurately classifying patients. All three target variables were considered for both these algorithms.

Decision Tree Features	Target variables
Chestnut – Trial and Total drug dosages: cyclophosphamide, etoposide, vincristine, L-asparaginase, steroid, methotrexate, purine, anthracycline, and cytarabine	4 class outcome: continued remission, remission death, relapse/refractory disease leading to 2nd remission, relapse/refractory disease leading to death.
Oak – Trial, regimen, delayed intensifications, purine received, and steroid received	Rel/ref: relapse/refractory disease, no relapse/refractory disease.
Elm – Trial, delayed intensifications, purine received, and steroid received	Remission death: died in remission, did not die in remission.

Table 49. Summary of the three decision trees and three target variables employed.

The data were randomly split into training and test datasets at a ratio of 70:30 using the train_test_split command. K-fold cross validation was used (with 10 folds) when utilising decision trees to ensure the random split was a good representation of the data, and that the accuracy of the trees were similar in each fold. This was performed with the commands KFold and cross_val_score. Similarly, repeated stratified K-Fold cross validation [Section 2.5.2.1] with 10 folds and 3 repeats was performed when utilising random forest and XGBoost algorithms with the command RepeatedStratifiedKFold. Decision trees were made using the DecisionTreeClassifier command, random forests with the RandomForestClassifier command, and XGBoost with the command XGBClassifier from the xgboost package. The metrics used to assess all the algorithms were accuracy, precision, recall, and F1 score values, as well as plotting confusion matrices and ROC curves. The commands for these were from the metrics library and were called accuracy_score, classification_report, confusion_matrix, and RocCurveDisplay.

Within decision trees, hyperparameter tuning using the GridSearchCV command [Section 2.5.2.2.2] and cost complexity pruning [Section 2.5.2.2.1] were both employed as methods to produce the most accurate decision tree that was also interpretable. GridSearchCV sought an optimal maximum depth and maximum number of features to be considered in a decision tree. A tree with a max depth of 5 and a minimum sample leaf of 15 was also created with the goal of obtaining a tree that was interpretable and created classifications of clinically relevant and statistically valid group sizes useful for analysis. For random forests, an equivalent method of hyperparameter tuning to GridSearchCV was used which was called RandomizedSearchCV which sought for an optimal maximum depth and number of estimators.

The issue of imbalanced data was addressed by assigning weight classes or resampling the data for all of the algorithms. For decision trees and random forests, balanced weight classes were assigned with the inbuilt option of class_weight, whilst scale_pos_weight was the inbuilt option for XGBoost to apply more weight to the minority class. Undersampling and oversampling were performed (as well as a combination of both) using the commands NearMiss (Version 1) and SMOTE respectively, both from the imbalanced-learn package. NearMiss was used with 5 nearest neighbours specified, whilst SMOTE used the 'auto' sampling strategy when applied to the data alone, and with a distribution of 486 patients (the original number of patients in the majority class) to each class when applied in conjunction with undersampling. T-SNE plots were used to assess the existence of clusters within the data, as well as to determine if the resampling produced an accurate representation of the data regarding those clusters. T-SNE plots were created using the TSNE command and plotted using plotly. An adaption of leave-one-out cross validation was used with random forest algorithms to attempt to include more events of interest in the training dataset, as an alternative methods to resampling to address class imbalance. For the random forests with relapse/ refractory disease as the target variable, 5 patients with an event were "left" in the test dataset due to a larger number of events. With remission death as the target variable, only one patient was "left" in the test dataset. The metrics reported for the leave-one-out cross validation are from the training dataset as opposed to the test dataset since the number of events left in the test dataset were too small to draw any meaningful conclusions from. However, the performance on the model on the test data was still assessed.

All machine learning was performed in Python using Jupyter Notebook version 6.4.8 in which a virtual environment was created to ensure consistent package versions were used every time. All of the above commands were in the scikit-learn package unless specified. The primary packages used throughout this project were: scikit-learn (version 1.4.2), imbalanced-learn (version 0.12.3), xgboost (version 2.0.3), pandas (version 1.4.2), plotly (version 5.22.0), matplotlib (version 3.5.1), and numpy (version 1.21.5). A random seed/state of 7 was used in each jupyter notebook as well as in the cross-validation, decision tree, random forest, and XGBoost commands to ensure reproducibility. A random state of 1 was used when splitting the data into training and test datasets and a random state of 42 was used for the t-SNE plots, also to ensure reproducibility.

5.4 Results

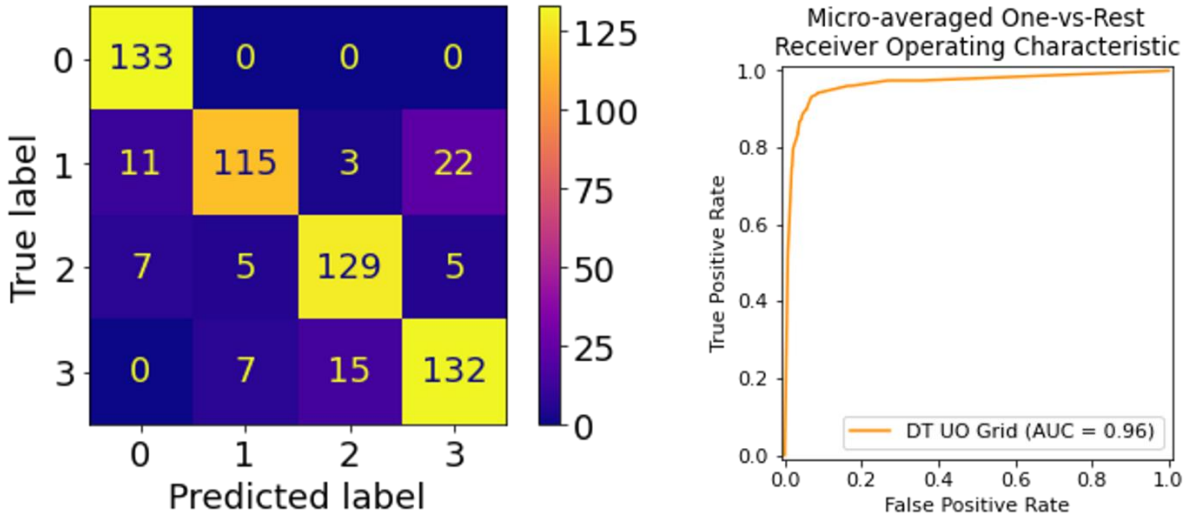
5.4.1 Decision Trees

The first machine learning algorithm utilised to classify good risk genetics patients by outcome was a decision tree. This was initially employed in the *ETV6*::*RUNX1* subset. The Chestnut, Oak, and Elm decision trees were built for all three target variables of interest. Cost complexity pruning, hyperparameter tuning using GridSearchCV, and pruning to a max depth of 5 with a minimum sample leaf of 15 were all explored. Balanced weight classes and resampling techniques were also assessed. 180 different decision tree model variations were considered, the results of which are presented in Supplementary Table 2. Trees were considered to be successful if an F1-score of ≥ 80% was achieved for each class.

5.4.1.1 ETV6::RUNX1

Within the *ETV6*::*RUNX1* dataset, neither Oak nor Elm were able to successfully classify patients in the 4 class outcome setting, with average F1-scores < 0.41 regardless of pruning, hyperparameter tuning, weighting classes, or resampling method [Supplementary Table 2]. The most accurate decision tree in this setting was the Chestnut tree, with a maximum depth of 7 and a maximum number of 7 features considered at each split as indicated by GridSearchCV, where the data had been both under- and over-sampled. This tree correctly classified all of the complete remission cases and >75% of the three other event classes, as shown by the recall values, and resulted in an averaged F1-score of 0.87 as well as an AUC of 0.96 [Figure 50]. However, the tree is too large to be easily interpretable, thus it is difficult to

draw meaningful conclusions from this tree. Therefore, trees with indicator variables as the target variable were explored, namely remission death and relapse/ refractory disease.

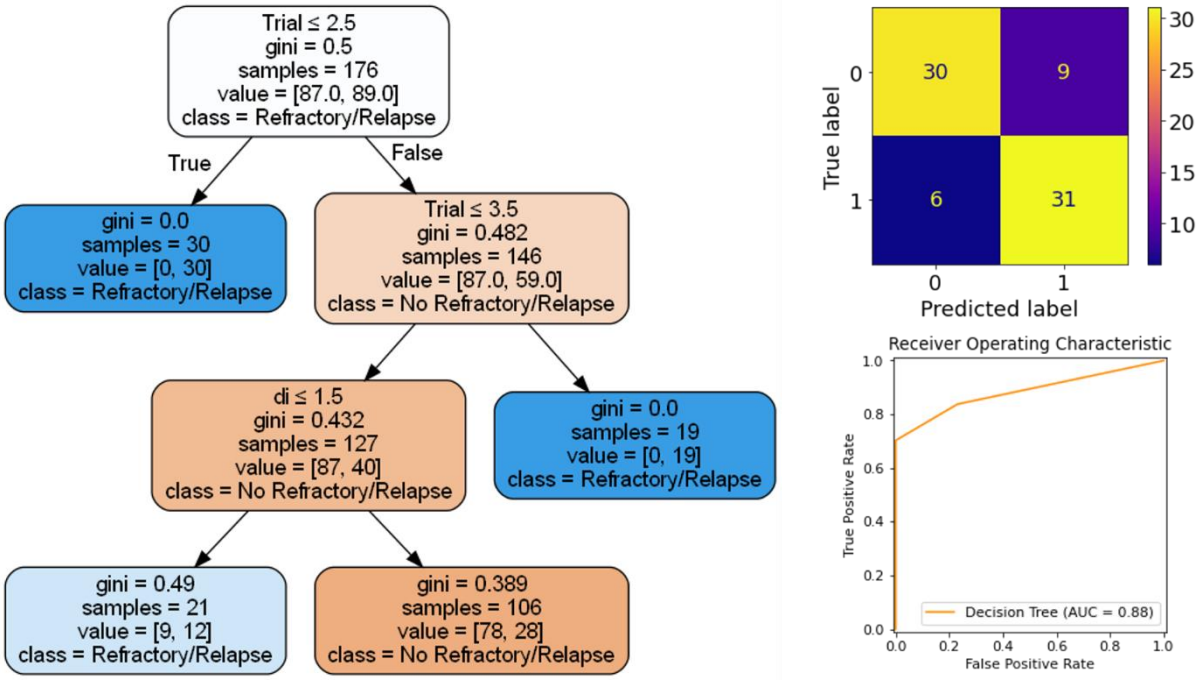


	Precision	Recall	F1-score	Accuracy
Complete Remission (= 0)	0.88	1	0.94	0.87
Remission Death (= 1)	0.91	0.76	0.83	
Relapse/ Refractory 2nd Rem (= 2)	0.88	0.88	0.88	
Relapse/ Refractory Death (= 3)	0.83	0.86	0.84	
Averaged metrics	0.87	0.88	0.87	

Figure 50. Metrics of the Chestnut decision tree with a max depth of 7 and a maximum number of features of 7, created with under- and over-sampled data. DT: decision tree, UO: under- and over-sampling, Grid: GridSearchCV, AUC = area under the curve.

With relapse/ refractory disease as the target variable, the Oak and Elm trees performed better, however, the Oak trees were still unable to accurately classify >75% of patients in the rel/ref group regardless of techniques applied [Supplementary Table 2]. The optimal Elm tree was a tree with a depth of 3 and 4 features where undersampling had been applied to the data. The tree had an AUC of 0.88 and correctly classified 84% of the rel/ref group and 77% of the non-rel/ref group [Figure 51]. The only two features used to classify patients were trial and number of delayed intensifications. All patients classified as non-rel/ref were patients on trials UKALL2003 and UKALL2011 who had 2 or more delayed intensifications. As all patients on UKALL2011 had 1 delayed intensification, this means the only patients the tree had classified as the non-rel/ref group were on UKALL2003 and had 2 delayed intensifications.

This is unsurprising given the fact that UKALL2003 patients had the best relapse rates of all the trials as shown in Chapter 3.

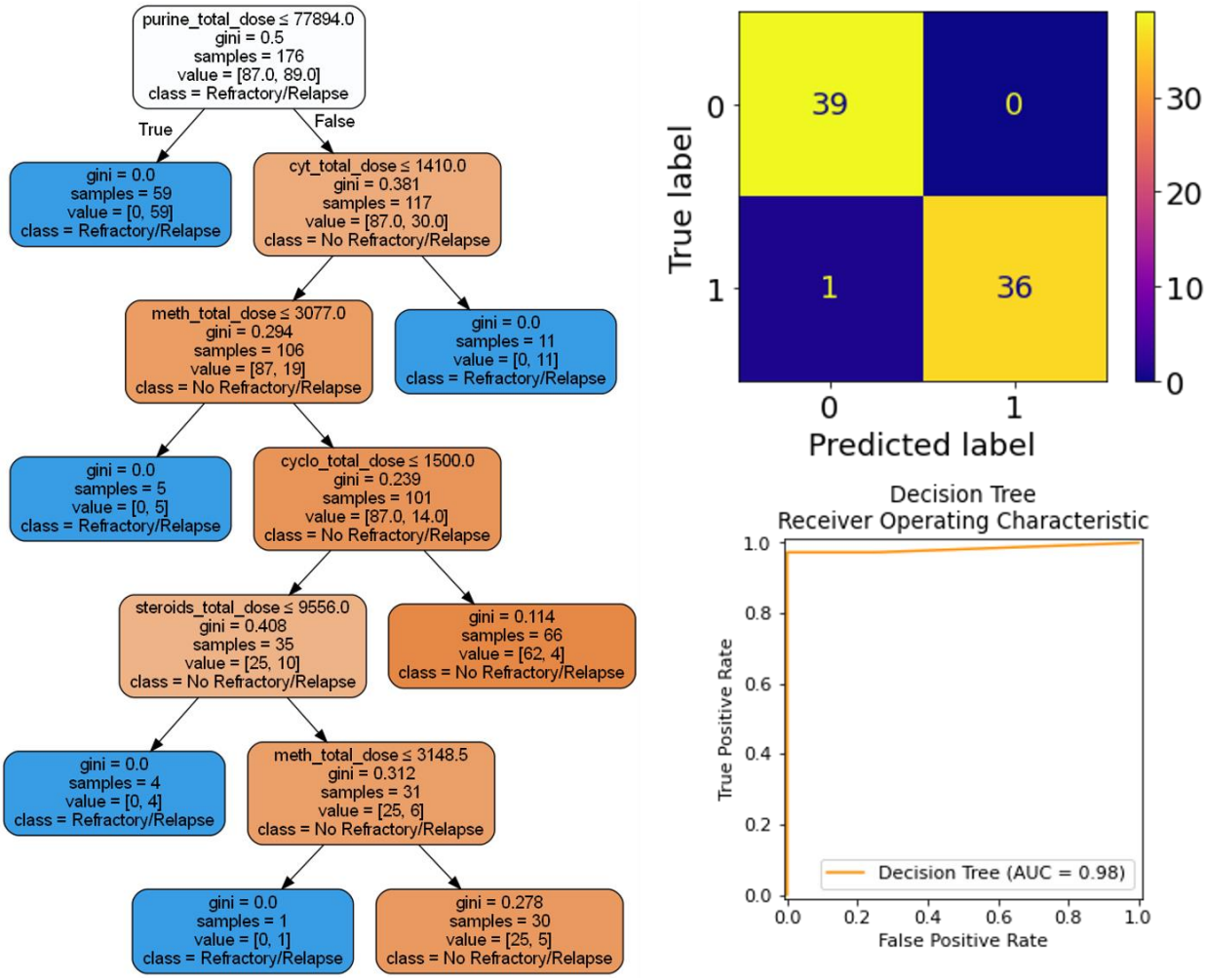


	Precision	Recall	F1-score	Accuracy
No Relapse/ Refractory Disease (= 0)	0.83	0.77	0.8	0.8
Relapse/ Refractory Disease (= 1)	0.78	0.84	0.81	
Averaged metrics	0.8	0.8	0.8	

Figure 51. Elm decision tree with a max depth of 3 and a maximum number of features of 4, created with undersampled data and the accompanying metrics. Trial was converted to a numerical variable with the following coding: UKALLXI92 = 0, UKALL97 = 1, UKALL97/99 = 2, UKALL2003 = 3, and UKALL2011 = 4. Thus, in the tree, ≤ 2.5 refers to trials UKALL97/99 or earlier. Delayed intensification is coded by number of delayed intensifications. Value refers to the number of training sample who are classified to each class, no relapse/refractory and relapse/refractory at each node. DI: delayed intensification, AUC: area under the curve.

The Chestnut tree was once again successful, with an average F1-score of 99% in the rel/ref setting when undersampling was performed, a max depth = 6, and max features = 4 were used. Only one patient was misclassified by the tree which was a relapse/refractory patient. Purine dose was the root feature of the tree, with cytarabine, methotrexate, cyclophosphamide, and steroid dose also included. The tree and the corresponding metrics are shown in Figure 52. Of note is that the optimal under- and over-sampled chestnut tree

also performed well in this setting with 100% classification accuracy in the non-rel/ref group and a correct classification rate of 91% in the rel/ref group [Supplementary Table 2]. As the decision trees seemed to more accurately classify patients with a binary target variable, remission death was also considered in this fashion.

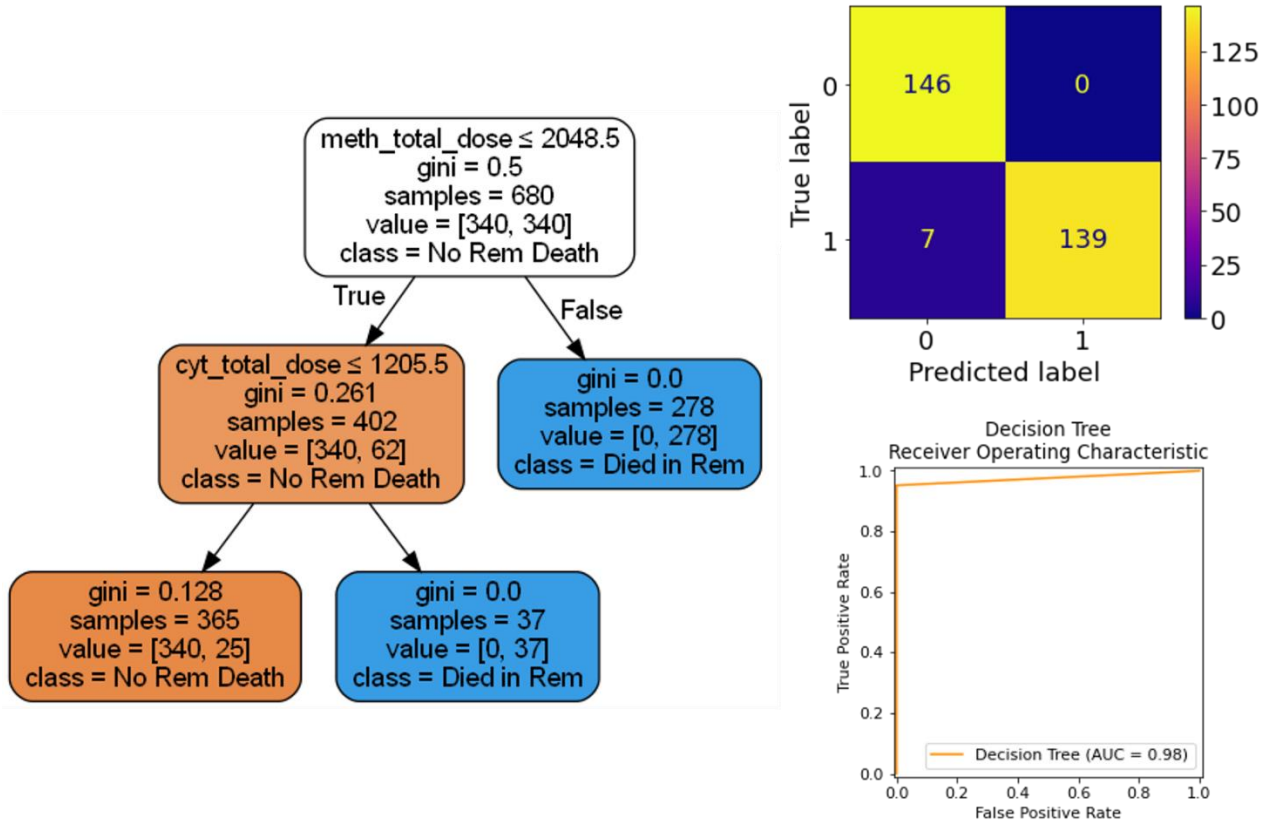


	Precision	Recall	F1-score	Accuracy
No Relapse/ Refractory Disease (= 0)	0.97	1	0.99	0.99
Relapse/ Refractory Disease (= 1)	1	0.97	0.99	
Averaged metrics	0.99	0.99	0.99	

Figure 52. Chestnut decision tree with a max depth of 6 and a maximum number of features of 4, created with undersampled data and the accompanying metrics. Cyt: cytarabine, meth: methotrexate, cyclo: cyclophosphamide, AUC: area under the curve.

The Chestnut tree was the only tree able to classify >75% of patients correctly when the target variable was remission death. The optimal Chestnut tree was created with under- and over-

sampled data. The tree did not need pruning, and hyperparameter tuning revealed that the full tree was the most optimal [Figure 53]. The tree had both an AUC and average F1-score of 0.98 and correctly classified 95% of the remission death patients and 100% of the other group. Methotrexate dose was the root feature in the tree with cytarabine dose being the only other feature.



	Precision	Recall	F1-score	Accuracy
No remission death (= 0)	0.95	1	0.98	0.98
Remission death (= 1)	1	0.95	0.98	
Averaged metrics	0.98	0.98	0.98	

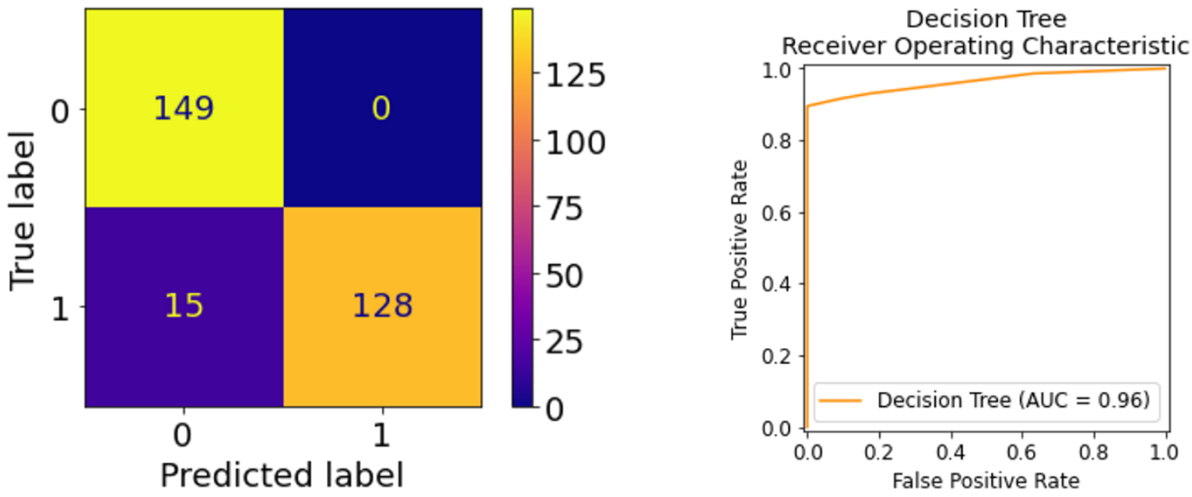
Figure 53. Chestnut decision tree, created with under- and over-sampled data and the accompanying metrics. Cyt: cytarabine, meth: methotrexate, AUC: area under the curve.

As under- and over-sampling was a successful method of handling imbalanced data for each target variable in *ETV6*::*RUNX1*, this was the only resampling technique considered in the high hyperdiploidy subgroup. All three trees were once again considered, as were the three target variables. The trees without any technique to resolve the class imbalance were once again unsuccessful and thus not presented.

5.4.1.2 High hyperdiploidy

Within the high hyperdiploidy dataset, where under- and over-sampling was performed on the data, neither Oak nor Elm were able to successfully classify >75% of patients in each outcome for any of the target variables. Chestnut was successful when relapse/ refractory disease and remission death were the target variables, but was unsuccessful when considering the 4 class outcome as the target, as only 45% of the rel/ref leading to 2nd remission group were correctly classified.

The optimal tree for classifying high hyperdiploidy patients by relapse/ refractory outcome was a Chestnut tree with a max depth of 6 and a maximum number of features of 3 considered at each split. This tree was able to correctly classify 100% of the non-rel/ref group and 90% of the rel/ref group resulting in an averaged F1-score and an AUC of 0.96 [Figure 54]. As is evidenced by Figure 55, the decision tree is more complex than many of the trees developed in the *ETV6*::*RUNX1* dataset and is therefore not as interpretable.



	Precision	Recall	F1-score	Accuracy
No Relapse/ Refractory Disease (= 0)	0.91	1	0.95	0.95
Relapse/ Refractory Disease (= 1)	1	0.9	0.94	
Averaged metrics	0.95	0.95	0.96	

Figure 54. Metrics of a chestnut decision tree with a max depth of 6 and a maximum number of features of 3, created with under- and over-sampled data. AUC: area under the curve

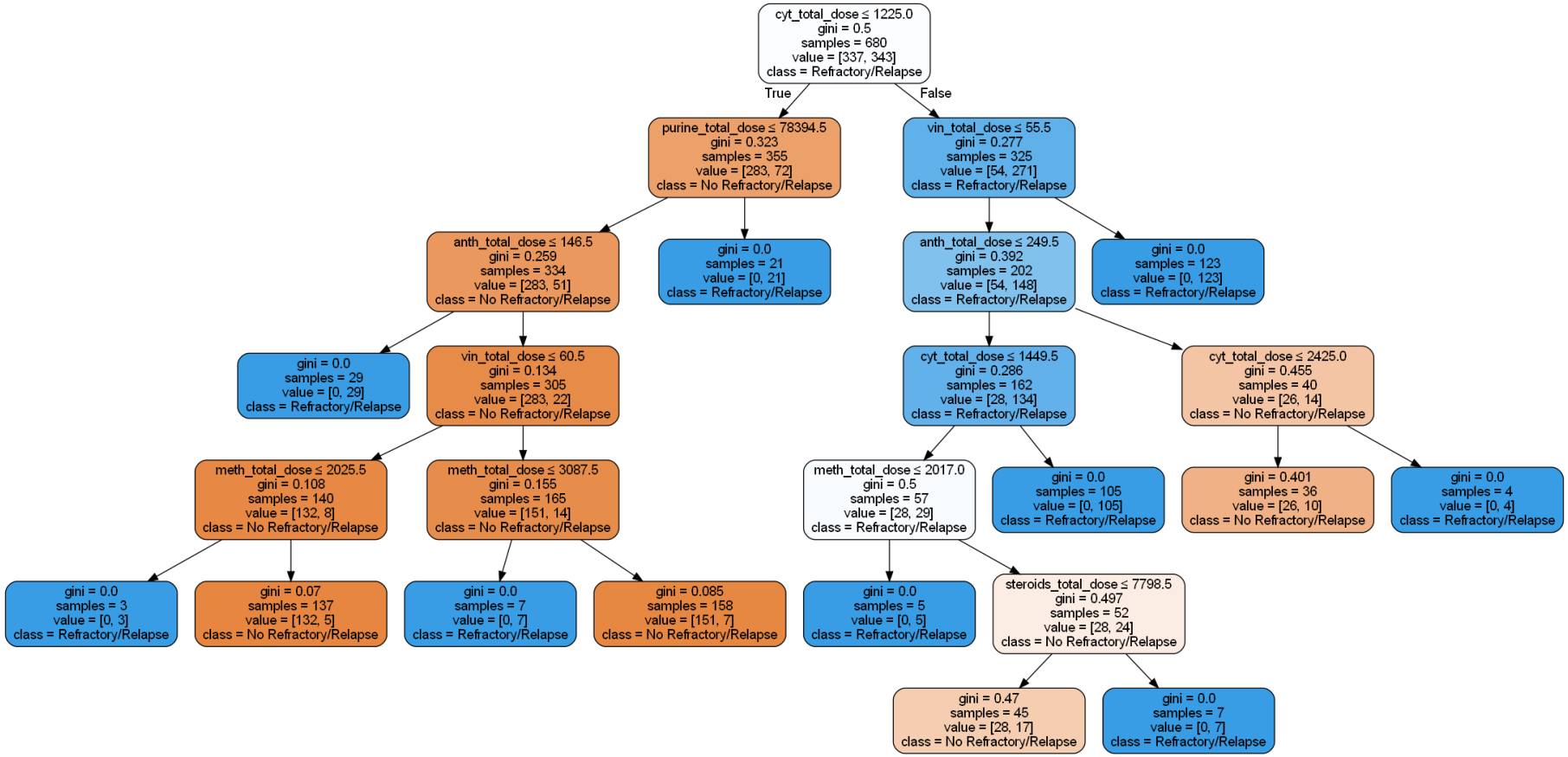
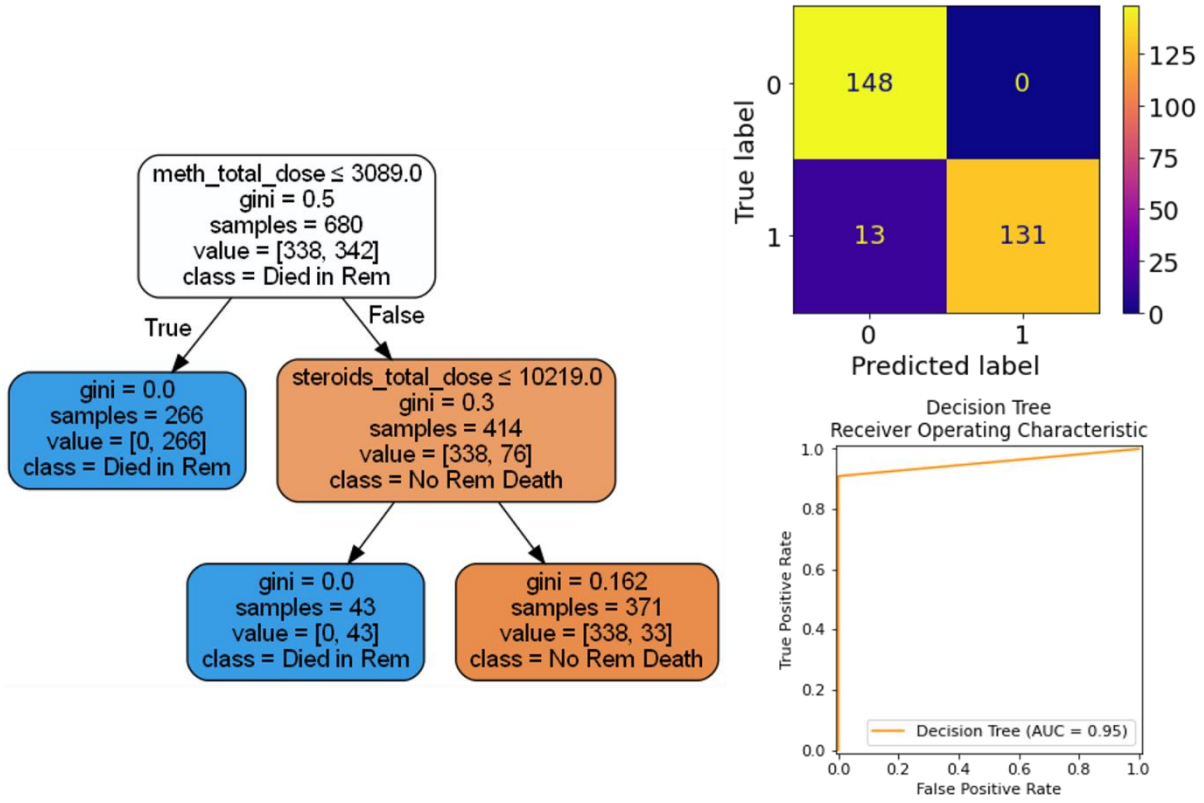


Figure 55. Chestnut decision tree with a max depth of 6 and a maximum number of features of 3, created with under- and over-sampled data.

When using remission death as the target variable, the optimal tree was the full Chestnut tree. The root feature for this tree was methotrexate dose whilst steroid dose was also used to classify patients. The tree was able to accurately classify every non-remission death patient, as well as 91% of the remission deaths. The tree had an AUC of 0.95 and an average F1-score of 0.96 [Figure 56].



	Precision	Recall	F1-score	Accuracy
No remission death (= 0)	0.92	1	0.96	0.96
Remission death (= 1)	1	0.91	0.95	
Averaged metrics	0.96	0.95	0.96	

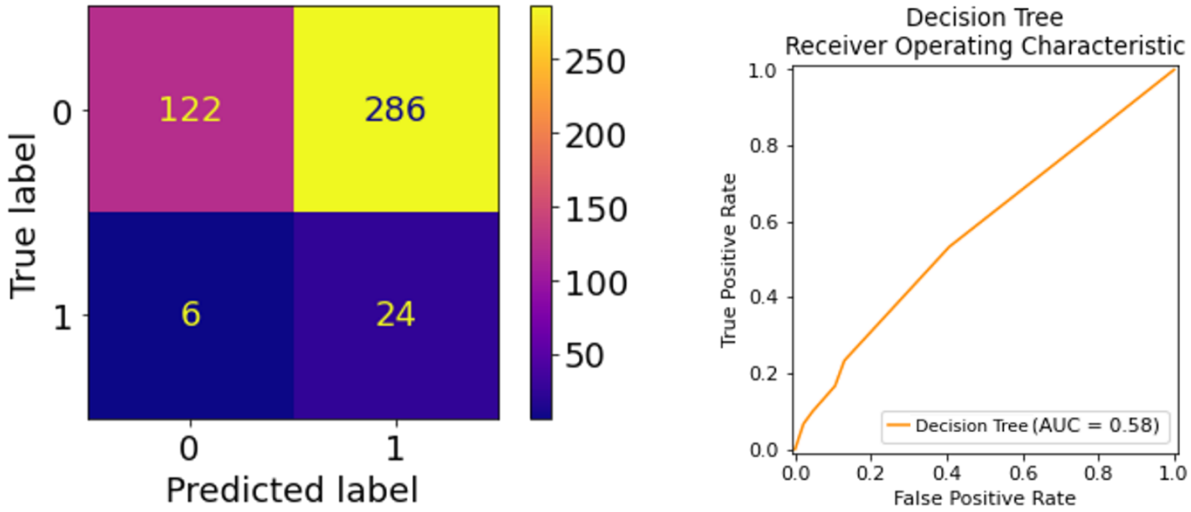
Figure 56. Chestnut decision tree, created with under- and over-sampled data and the accompanying metrics. Meth: methotrexate, AUC: area under the curve.

5.4.1.3 Testing the trees in the original data

As all the successful trees had been produced with resampled data, these trees were tested in the original data to see if the splits found truly represented groups with different outcomes. The results showed that resampled data did not honour the boundaries of the groups found in the actual data, as in order to correctly classify the majority of events of interest (e.g.

remission deaths) the tree also classified most cases without the event of interest that way (e.g. non-remission deaths). This occurred in both *ETV6*::*RUNX1* and high hyperdiploidy data and an example is presented for a tree in each subset below.

The Elm decision tree, aiming to classify relapse/refractory disease in *ETV6*::*RUNX1* patients, was able to successfully classify 80% of the rel/ref patients in the original data, however, this was at the expense of misclassifying 70% of the non-rel/ref cases where it had only misclassified 23% in the resampled data. This resulted in an AUC of 0.58 which indicates that this tree is only a slight improvement upon a random classification of these patients [Figure 57].

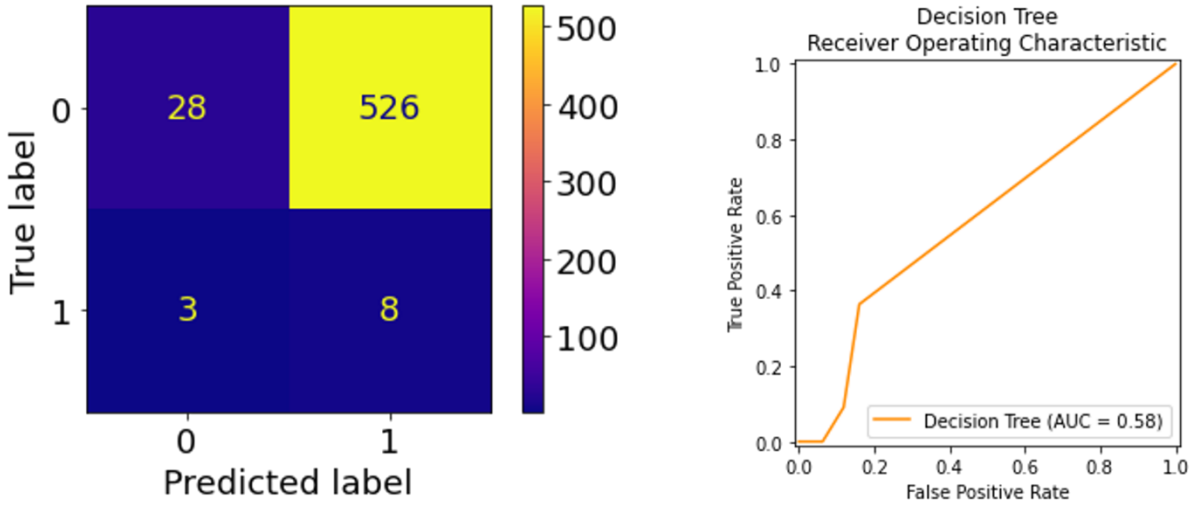


	Precision	Recall	F1-score	Accuracy
No Relapse/ Refractory Disease (= 0)	0.95	0.3	0.46	0.33
Relapse/ Refractory Disease (= 1)	0.08	0.8	0.14	
Averaged metrics	0.52	0.55	0.3	

Figure 57. Metrics of the Elm decision tree with a max depth of 7 and a maximum number of features of 7, created with under- and over-sampled data applied to the original *ETV6*::*RUNX1* data. AUC: area under the curve.

Similarly, the Chestnut decision tree aiming to classify remission deaths in high hyperdiploidy patients incorrectly classified 95% of non-remission deaths in order to be able to classify 73% of the remission death patients correctly. This is a stark contrast from the performance in the resampled data, in which 100% of non-remission deaths and 91% of remission deaths were

correctly classified by the tree. This poor performance, again, resulted in an AUC just above 0.5 showing that this tree makes little improvement upon a random classification [Figure 58].



	Precision	Recall	F1-score	Accuracy
No remission death (= 0)	0.9	0.05	0.1	0.06
Remission death (= 1)	0.01	0.73	0.03	
Averaged metrics	0.46	0.39	0.06	

Figure 58. Metrics of the Chestnut decision tree created with under- and over-sampled data applied to the original high hyperdiploidy data. AUC: area under the curve.

5.4.2 t-SNE

In order to determine if there were indeed clusters in the data that the machine learning algorithms could identify, as well as to confirm that the resampled data did not retain the boundaries between the classes present in the original data, t-SNE visualisation [Section 2.4.1] was performed both in *ETV6*::*RUNX1* and high hyperdiploidy, with the drug dosages and trial as features since the Chestnut tree seemed best at classifying patients. Trial was kept as a numeric variable for the t-SNE visualisation in order to remain consistent with the machine learning.

For *ETV6*::*RUNX1* patients, it was clear that there were distinct clusters within the data, however, these clusters were not based on outcome as is clear from Figure 59. Instead, rel/ref cases seem to be clustered very closely with non-rel/ref cases, likely explaining the decision trees' inability to find splits that classify patients based on outcome. The perplexity is a tuneable parameter which is assigned to balance the attention between local and global

aspects of the data. It is an estimate of the number of close neighbours each data point will have. A perplexity of 45 was set in this study as this was deemed to be the most effective within the *ETV6*::*RUNX1* cohort. Perplexity optimisation was performed by assessing the Kullback-Leibler divergence at different perplexity values between a range of 5 and 55 with the goal of choosing the perplexity value at which the divergence begins to stabilise.

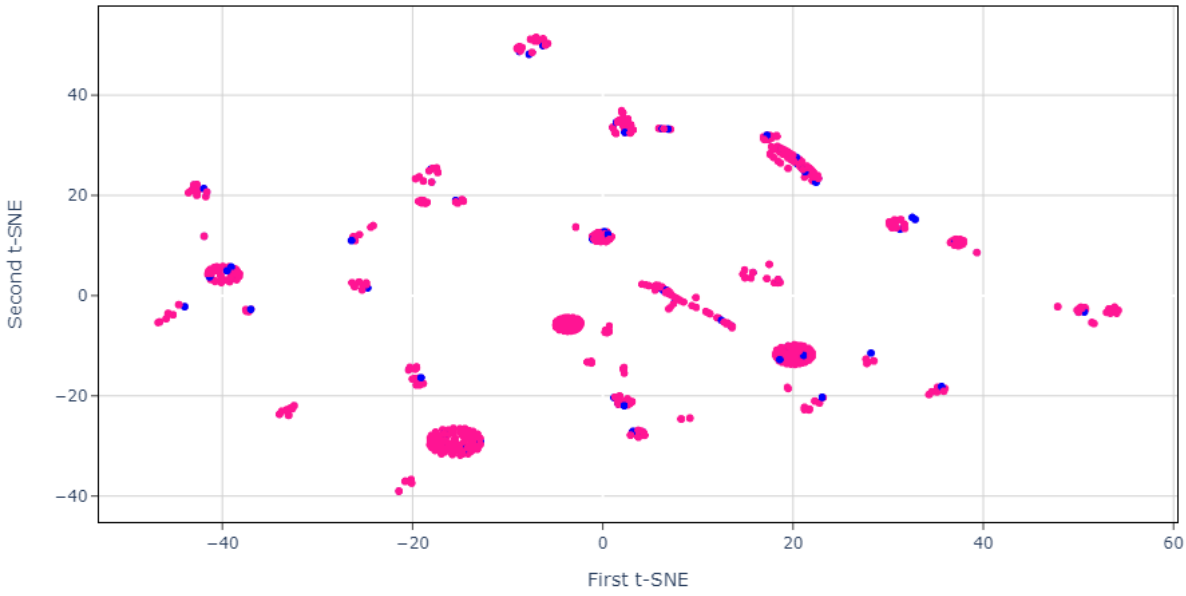


Figure 59. t-SNE visualisation of the original *ETV6*::*RUNX1* data coloured by relapse/refractory disease. Blue denotes the relapse refractory disease cases and pink denotes non-relapse/ refractory disease cases.

The visualisation of the resampled *ETV6*::*RUNX1* data differs dramatically from the original data, with distinct clusters of rel/ref cases and largely distinct clusters of non-rel/ref cases [Figure 60]. This explains the decision trees' ability to classify patients based on the features in the chestnut tree when the data were resampled.

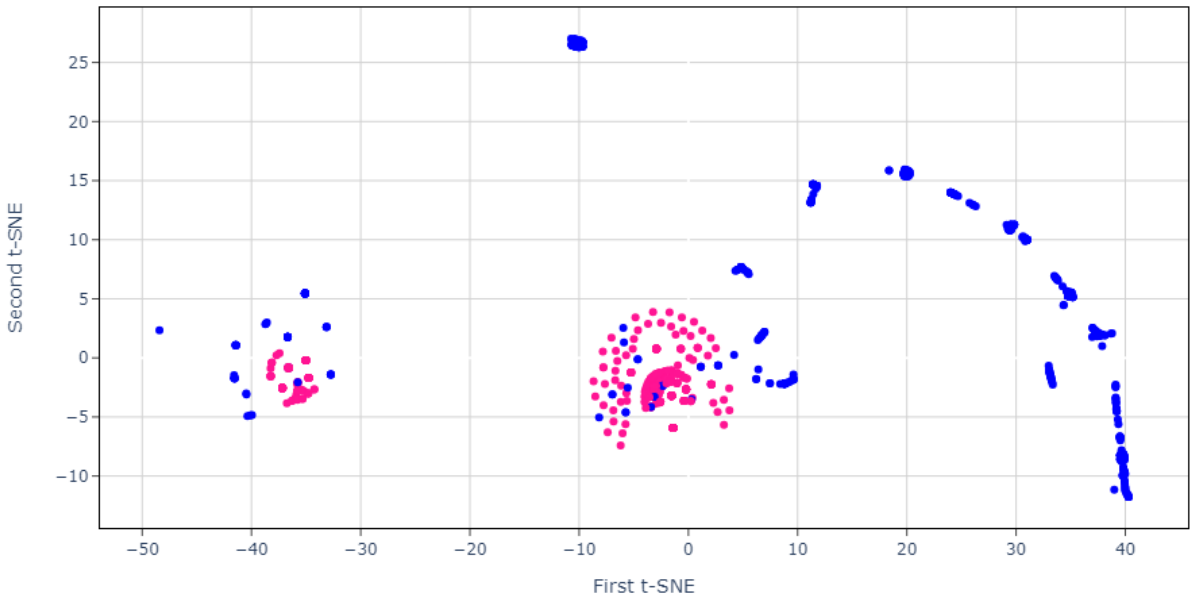


Figure 60. t-SNE visualisation of the under- and over-sampled *ETV6*::*RUNX1* data coloured by relapse/refractory disease. Blue denotes the relapse refractory disease cases and pink denotes the non-relapse/ refractory disease cases.

Within the high hyperdiploidy dataset, there were also no distinct clusters based on outcome as rel/ref cases grouped with non-rel/ref cases in the original data, although clusters did exist [Figure 61].

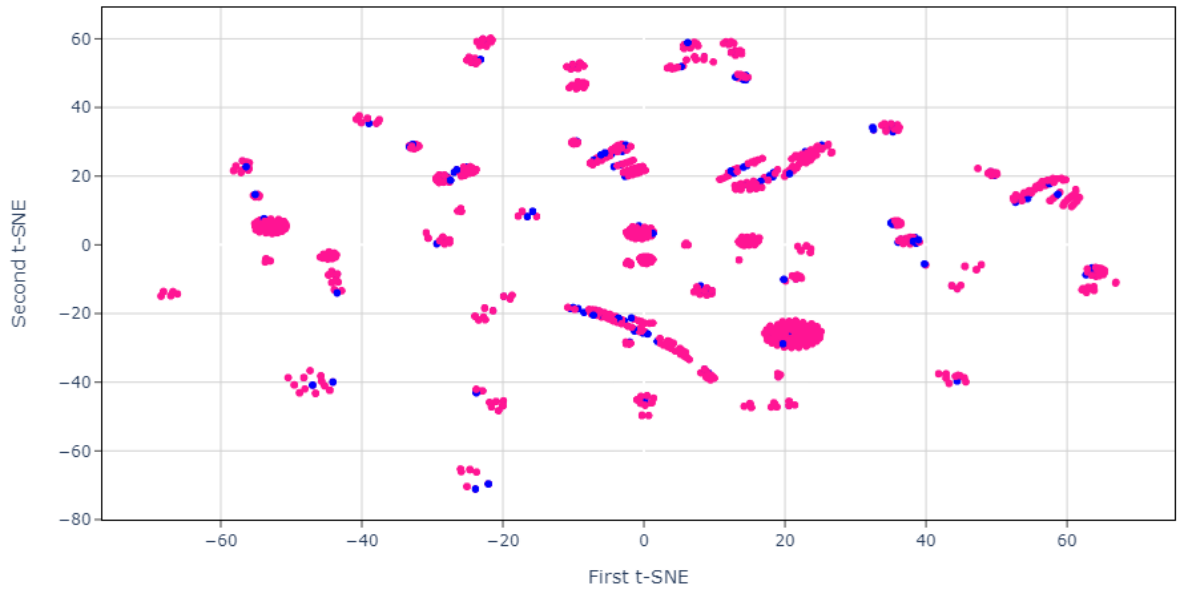


Figure 61. t-SNE visualisation of the original high hyperdiploidy data coloured by relapse/refractory disease. Blue denotes the relapse refractory disease cases and pink denotes the non-relapse/ refractory disease cases.

This was, again, completely different in the resampled data where there were distinct clusters of rel/ref cases and largely distinct clusters of non-rel/ref cases [Figure 62]. This shows that the resampled data were not a true representation of the original data in either genetic subgroup, elucidating the lack of generalisability of the decision trees created in the resampled data.

The resampled data likely show very different clusters to the original data as the data points that were neighbours of the events of interest have been removed, and new synthetic minority samples have been included created artificial groups of majority and minority samples that are completely distinct from one another. Thus when the trees are employed in the original data, the samples from the majority group that were clustered near to the samples that had the event of interest were misclassified as events of interest. From these findings, it is plausible that these features cannot be used alone to predict good risk genetic patient outcomes. This is possibly because the values of these features are too similar amongst the classes, perhaps because not enough continuous variables were considered, and thus no viable splits can be found. However, as this data is complex, it is also feasible that

there are clusters which are not apparent and that the decision tree algorithm was simply unable to identify splits to accurately predict these groups.

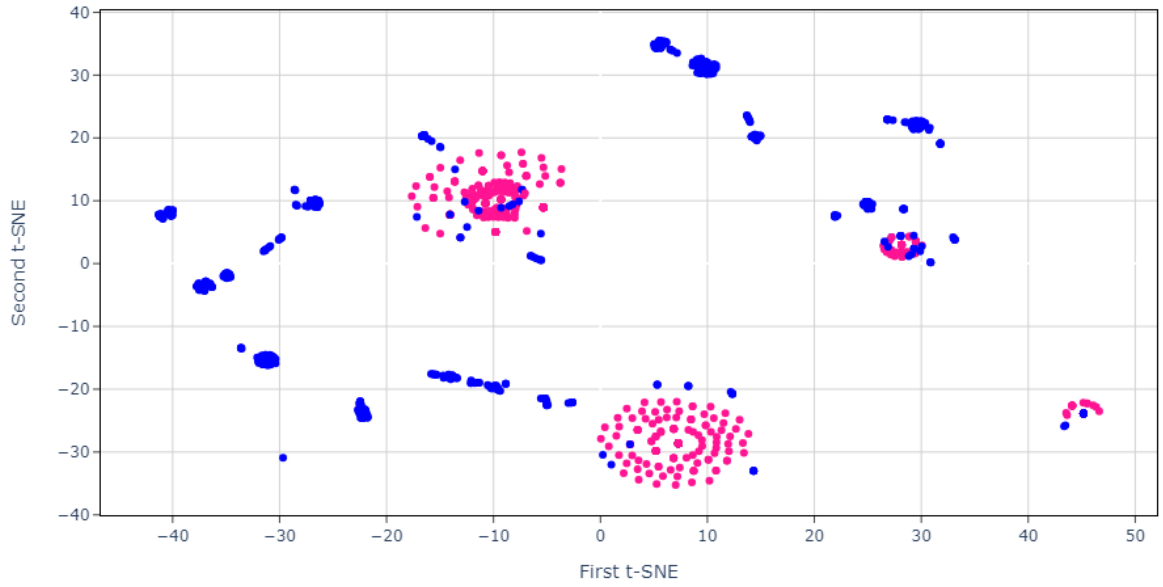


Figure 62. t-SNE visualisation of the under- and over-sampled high hyperdiploidy data coloured by relapse/refractory disease. Blue denotes the relapse refractory disease cases and pink denotes the non-relapse/ refractory disease cases.

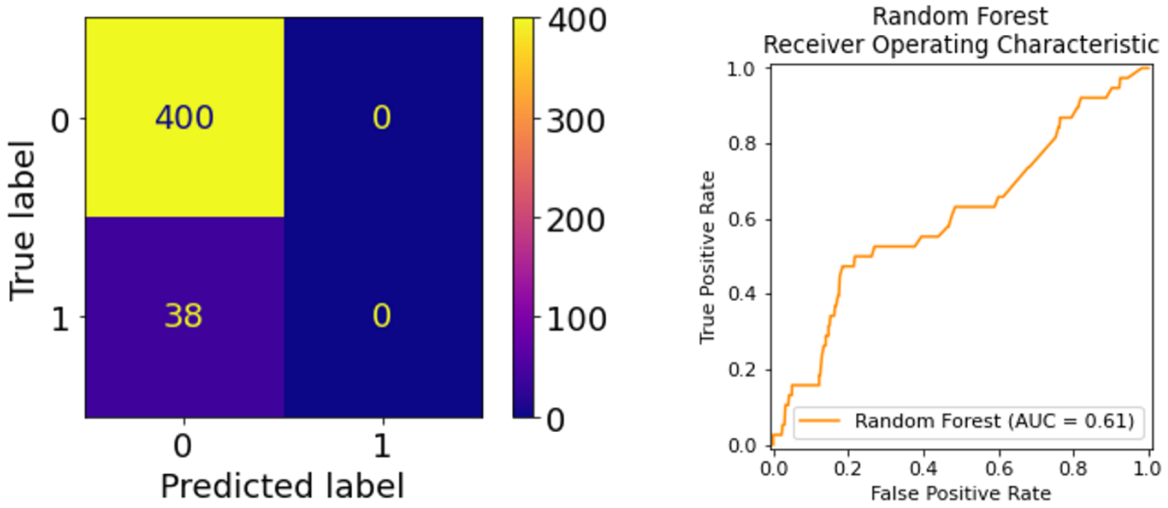
5.4.3 Bagging

As it was clear that the decision tree algorithm was not able to accurately classify patients based on outcome, a bagging technique was explored – the random forest algorithm. This algorithm is generally more accurate than the decision tree as it has a higher predictive power whilst also being more robust and reducing overfitting. The Chestnut tree was the only one taken forward for bagging as these were the only features that were optimal for the decision trees to predict patient outcomes [Supplementary Table 2]. All three target variables were considered.

5.4.3.1 ETV6::RUNX1

It is clear from Figure 63 that the random forest algorithm was not an improvement over the decision tree in *ETV6*::*RUNX1* data with relapse/ refractory disease as the target variable. It simply classified all patients as non-relapse/ refractory disease as this still resulted in a 91% accuracy overall due to the imbalanced classes. This was also the case in the 4 class outcome

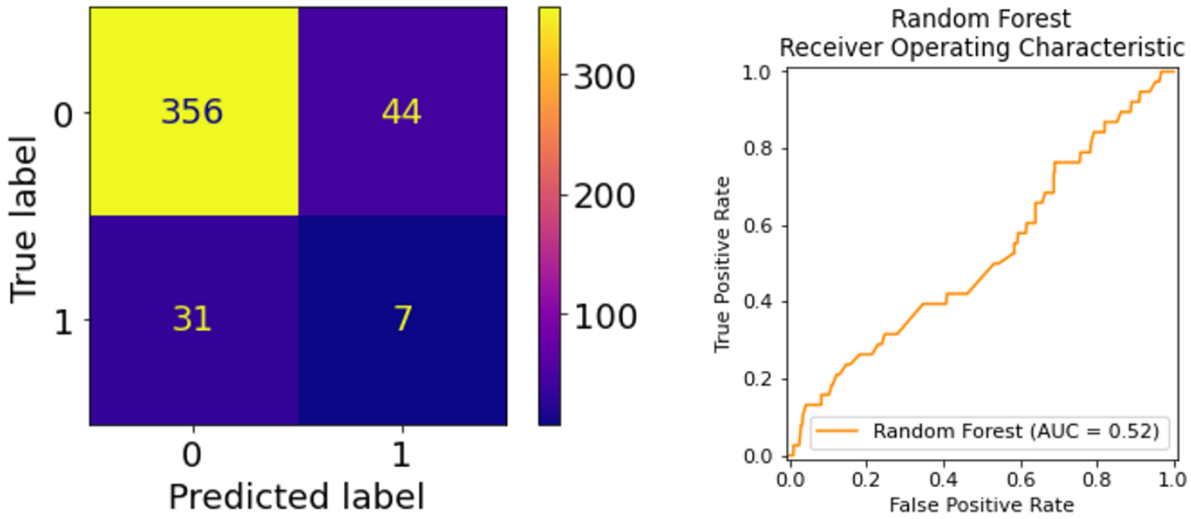
and remission death settings. Thus, the stipulation of balanced weight classes was added into the algorithm in an effort to ensure the algorithm focused on correctly classifying every group.



	Precision	Recall	F1-score	Accuracy
No Relapse/ Refractory Disease (= 0)	0.91	1	0.95	0.91
Relapse/ Refractory Disease (= 1)	0	0	0	
Averaged metrics	0.46	0.5	0.48	

Figure 63. Metrics of a Chestnut random forest created in *ETV6*::*RUNX1* data with a max depth of 3 and 84 estimators where relapse/ refractory disease is the target variable. AUC: area under the curve.

Whilst the balanced weight classes did result in some relapse/ refractory disease cases being correctly classified, 82% were still misclassified and this minor improvement was at the detriment of 11% of non-rel/ref cases that were now misclassified [Figure 64]. This resulted in an average F1-score of 0.53 and an AUC of 0.52. This was also the case when the other two target variables were considered (<25% accuracy in the minority classes). Therefore it is clear that balanced weight classes were not able to resolve the issue of imbalanced data within the *ETV6*::*RUNX1* subgroup.

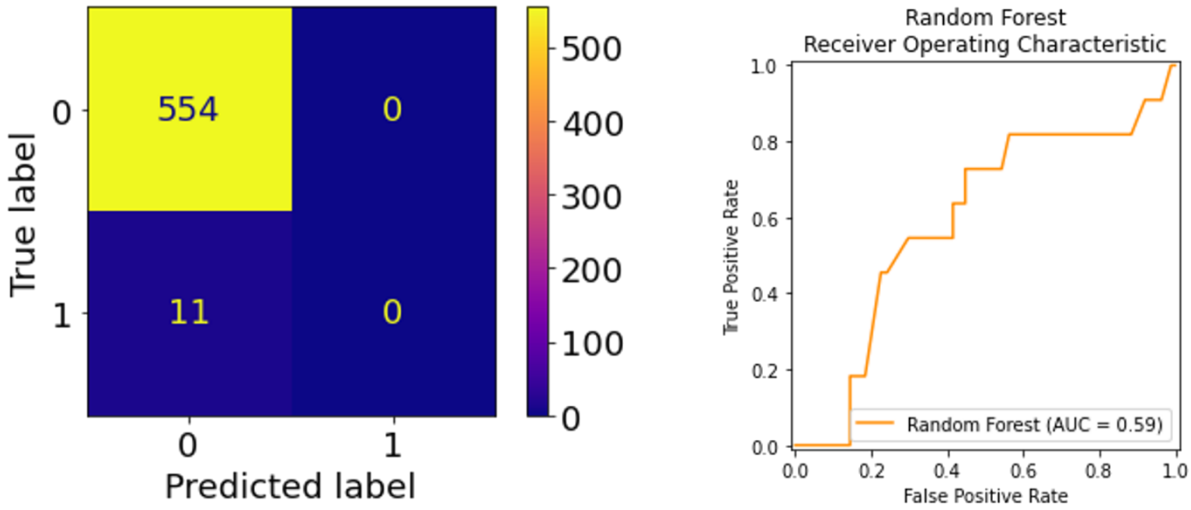


	Precision	Recall	F1-score	Accuracy
No Relapse/ Refractory Disease (= 0)	0.92	0.89	0.9	0.83
Relapse/ Refractory Disease (= 1)	0.14	0.18	0.16	
Averaged metrics	0.53	0.54	0.53	

Figure 64. Metrics of a Chestnut random forest created in *ETV6*::*RUNX1* data with a max depth of 13, 142 estimators, balanced class weights, and relapse/ refractory disease as the target variable. AUC: area under the curve.

5.4.3.2 High hyperdiploidy

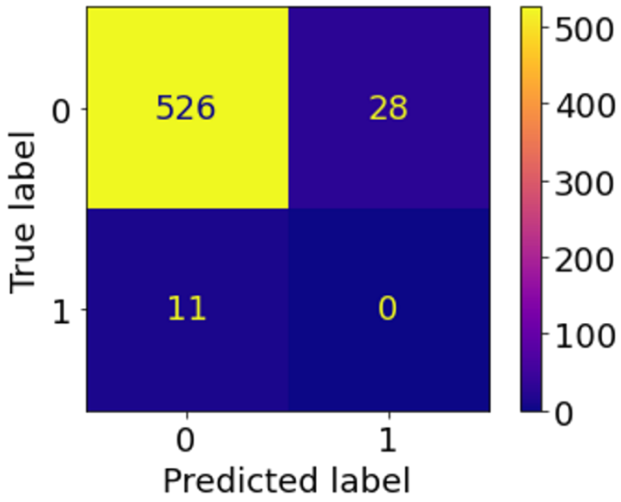
Within high hyperdiploidy, the random forest performed much the same and classified all patients as non-remission death, as this resulted in the best accuracy of 83% due to the imbalanced nature of the data [Figure 65]. The averaged F1-score was 0.5 and the AUC was 0.59 as one group was completely misclassified. Again, this occurred in both the relapse/ refractory disease and 4 class outcome settings.

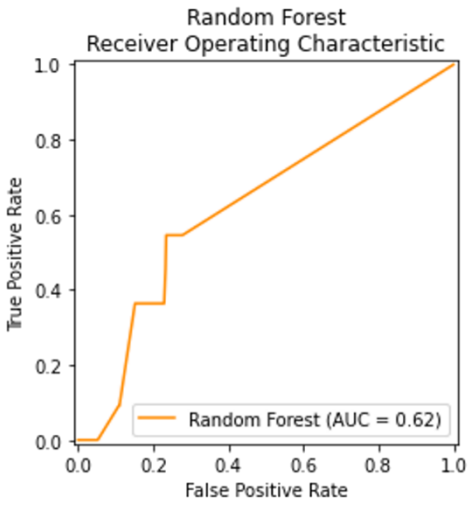


	Precision	Recall	F1-score	Accuracy
No remission death (= 0)	0.98	1	0.99	0.98
Remission death (= 1)	0	0	0	
Averaged metrics	0.49	0.5	0.5	

Figure 65. Metrics of a Chestnut random forest created in high hyperdiploidy data with a max depth of 4 and 261 estimators where remission death is the target variable. AUC: area under the curve.

The balanced class weights did not improve the misclassification of the remission death patients within the high hyperdiploidy subgroup and instead misclassified 5% of non-remission deaths. This resulted in a lower accuracy of 93% compared to the standard random forest and a lower average F1-score of 0.48 [Figure 66]. Similar occurrences happened within the rel/ref and 4 class outcome settings where there was marginal improvement in the minority groups classes (<20% accuracy in each minority class) and a decrease in classification accuracy of the majority class.





	Precision	Recall	F1-score	Accuracy
No remission death (= 0)	0.98	0.95	0.96	0.93
Remission death (= 1)	0	0	0	
Averaged metrics	0.49	0.47	0.48	

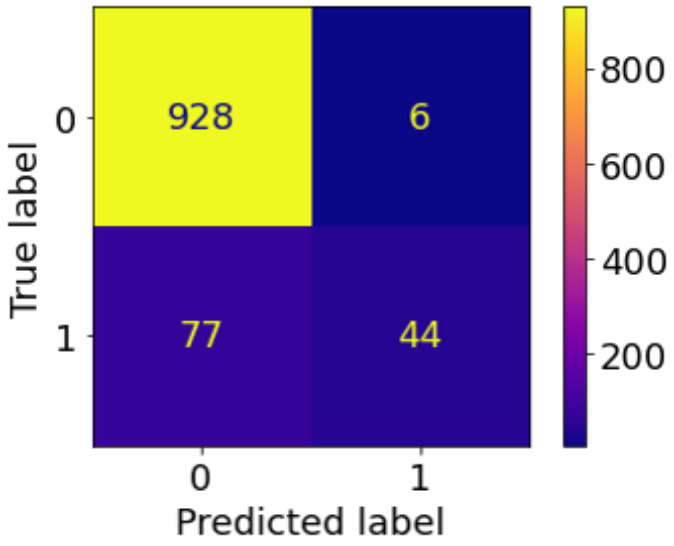
Figure 66. Metrics of a Chestnut random forest created in high hyperdiploidy data with a max depth of 16, 246 estimators, balanced class weights, and remission death as the target variable. AUC: area under the curve.

5.4.3.3 Leave-one-out cross-validation

As the t-SNE visualisation confirmed that resampling the data was not viable, an alternative method of handling the imbalanced data was investigated – a modification of leave-one-out cross validation. This was only performed with relapse/ refractory disease and remission death as the target variable since the goal of this approach is to have as many events as possible in the training dataset, and thus splitting the events into three groups was not logical. This cross-validation approach was performed 10-fold to ensure the results reported were not affected by variability in the data. The random forests were also subject to hyperparameter tuning.

The random forest employed on *ETV6*::*RUNX1* data with relapse/ refractory disease as the target variable was not able to correctly classify the minority group. When assessed in the training dataset, approximately two thirds of the rel/ref group were misclassified and the average F1-score was 74% [Figure 67]. The hyperparameter tuning resulted in a tree which classified all patients as non-rel/ref. These results were similar across the 10 folds. When

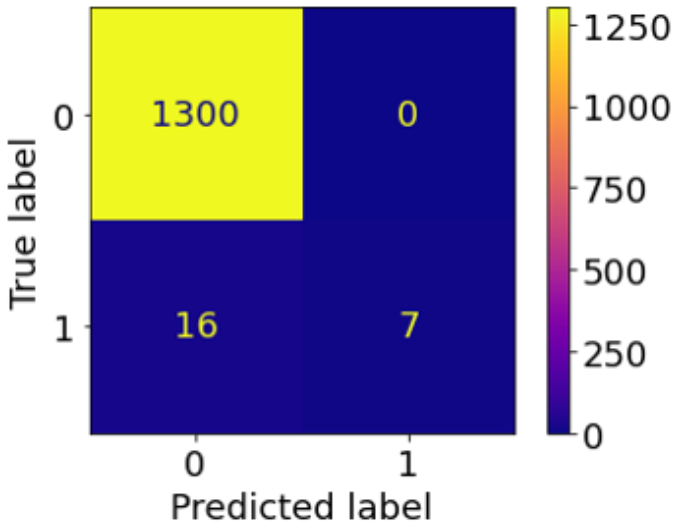
assessed in the test dataset, across the 10 folds, the random forest most frequently misclassified all five rel/ref patients with the exception of three times where one rel/ref patient was correctly classified.



	Precision	Recall	F1-score	Accuracy
No Relapse/ Refractory Disease (= 0)	0.92	0.99	0.96	0.92
Relapse/ Refractory Disease (= 1)	0.88	0.36	0.51	
Averaged metrics	0.9	0.68	0.74	

Figure 67. Metrics of a Chestnut random forest created in *ETV6*::*RUNX1* data using leave-five-out cross validation, with relapse/ refractory disease as the target variable, where the training data were used to create these metrics.

Within the high hyperdiploidy data, the random forest classified 100% of non-remission death patients correctly but only 30% of remission death patients [Figure 68]. This resulted in a very high accuracy (99%) but an average F1-score of 0.73. These findings were similar across the 10 folds. The random forest with hyperparameter tuning again classified all patients as non-remission death most frequently throughout the 10 folds. When assessing the performance of the model in the test data, the one remission death patient was misclassified 9 times out of 10 by the algorithm regardless of hyperparameter tuning.



	Precision	Recall	F1-score	Accuracy
No Remission death (= 0)	0.99	1	0.99	0.99
Remission death (= 1)	1	0.3	0.47	
Averaged metrics	0.99	0.65	0.73	

Figure 68. Metrics of a Chestnut random forest created in high hyperdiploidy data using leave-one-out cross validation, with remission death as the target variable, where the training data were used to create these metrics.

It is clear from these results that the bagging technique was unsuccessful as the algorithm consistently failed to correctly classify the minority class, as evidenced by the metrics presented in this section.

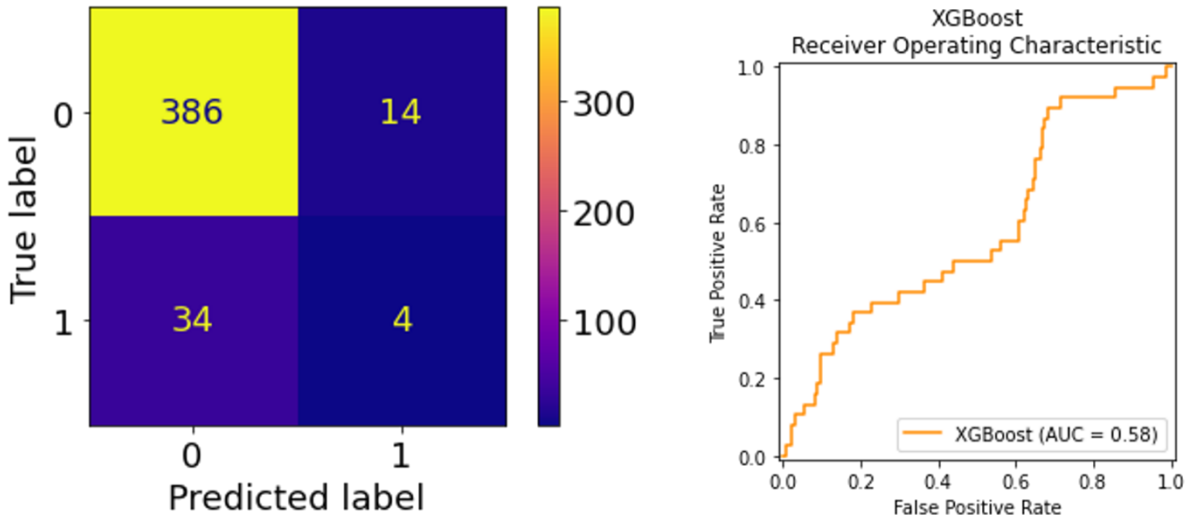
5.4.4 Boosting

As both the decision tree [Section 5.4.2] and random forest [Section 5.4.3] were not successful at classifying patients, a boosting algorithm was explored. Boosting has been proposed as a method for handling imbalanced data as it penalises incorrect classifications and assigns more weight to correct classifications. The algorithm XGBoost was employed in both the *ETV6*::*RUNX1* and high hyperdiploidy subgroups. All three target variables were once again considered.

5.4.4.1 ETV6::RUNX1

Within the *ETV6*::*RUNX1* dataset, the XGBoost algorithm with Chestnut features performed poorly for each target variable. The algorithm successfully classified 11% of relapse/ refractory disease patients and had an AUC of 0.58 [Figure 69]. Similar results were found in

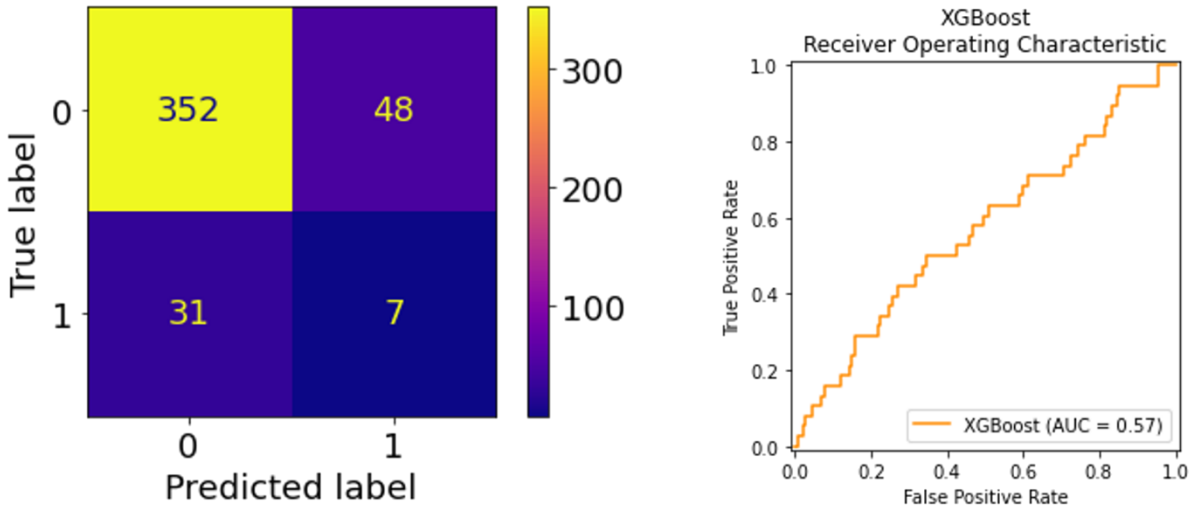
the other settings where only 17% of remission death patients were correctly classified, whilst each minority class had a recall of <0.18 in the 4 class setting.



	Precision	Recall	F1-score	Accuracy
No Relapse/ Refractory Disease (= 0)	0.92	0.96	0.94	0.89
Relapse/ Refractory Disease (= 1)	0.22	0.11	0.14	
Averaged metrics	0.57	0.54	0.54	

Figure 69. Metrics of an XGBoost algorithm created in *ETV6*::*RUNX1* data with Chestnut features, where relapse/ refractory disease is the target variable. AUC: area under the curve.

Since the algorithm was still misclassifying the minority class samples, a weighted XGBoost algorithm was utilised with a weighting of 1000 towards the minority class. This resulted in the correct classification of 3 additional relapse/ refractory patients at the detriment of misclassifying an additional 34 non-rel/ref cases. Furthermore, the recall of the rel/ref group was only 0.18 and the AUC for the algorithm was 0.57 [Figure 70]. This same occurrence happened with the other two target variables with recall values <0.18 for each minority class. Within the 4 class outcome setting, the weighted algorithm offered no improvement over the standard algorithm.

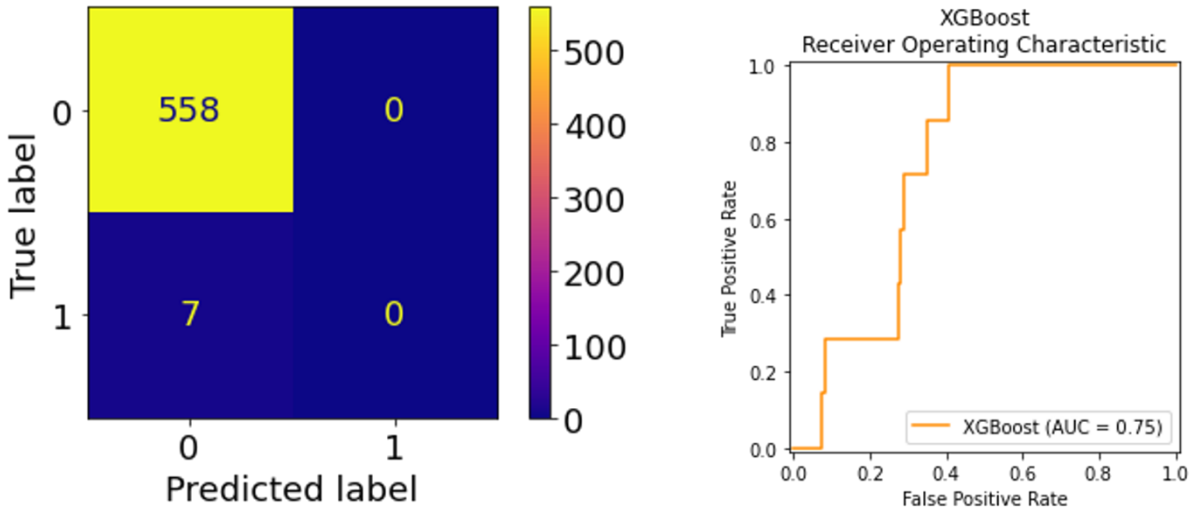


	Precision	Recall	F1-score	Accuracy
No Relapse/ Refractory Disease (= 0)	0.92	0.88	0.9	0.82
Relapse/ Refractory Disease (= 1)	0.13	0.18	0.15	
Averaged metrics	0.52	0.53	0.52	

Figure 70. Metrics of a weighted XGBoost algorithm created in *ETV6*::*RUNX1* data with Chestnut features, where relapse/ refractory disease is the target variable. AUC: area under the curve.

5.4.4.2 High hyperdiploidy

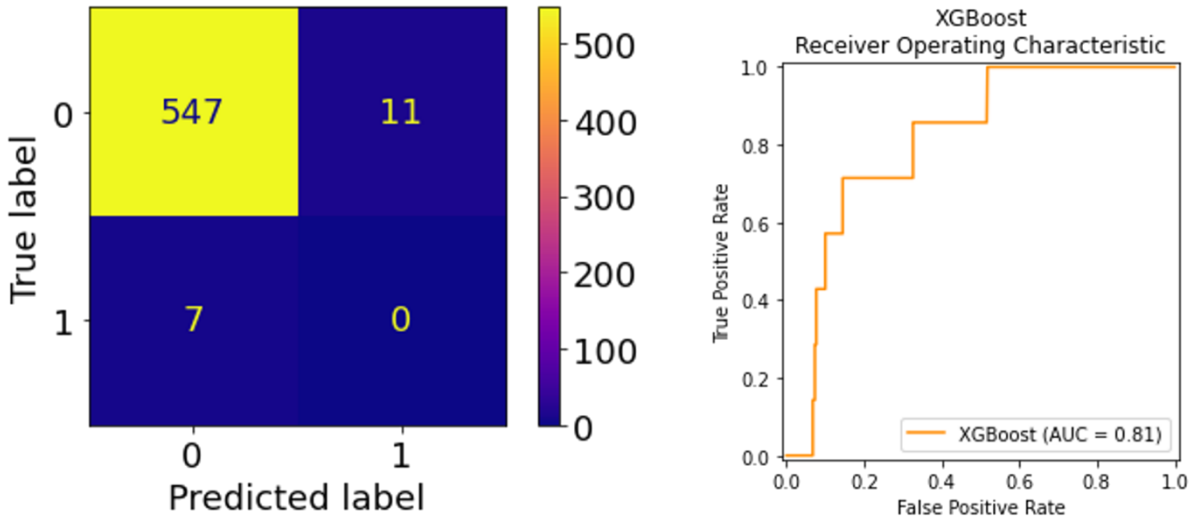
The XGBoost algorithm performed similarly within the high hyperdiploidy subset as it completely misclassified the remission death group by classifying all patients as non-rel/ref [Figure 71]. With relapse/ refractory disease as the target variable, 98% of the minority class patients were misclassified. For the 4 class outcome, only 6% of rel/ref patients leading to 2nd remission were able to be accurately classified, whilst the patients in the other two minority classes were all misclassified.



	Precision	Recall	F1-score	Accuracy
No remission death (= 0)	0.99	1	0.99	0.99
Remission death (= 1)	0	0	0	
Averaged metrics	0.49	0.5	0.5	

Figure 71. Metrics of an XGBoost algorithm created in high hyperdiploidy data with Chestnut features, where remission death is the target variable. AUC: area under the curve.

A weighted XGBoost algorithm was once again utilised with a weighting of 1000 towards the minority class. This did not improve the model as 100% of the rel/ref class group were still misclassified, whilst 2% of the non-rel/ref group were now misclassified [Figure 72]. Similarly, in the 4 class outcome, only 6% of rel/ref patients (regardless of outcome) were correctly classified whilst all of the remission death patients remained misclassified by the model. For relapse/ refractory as the target variable, 34% of minority class and 77% of the non-rel/ref patients were accurately classified.



	Precision	Recall	F1-score	Accuracy
No remission death (= 0)	0.99	0.98	0.98	0.97
Remission death (= 1)	0	0	0	
Averaged metrics	0.49	0.49	0.49	

Figure 72. Metrics of a weighted XGBoost algorithm created in high hyperdiploidy data with Chestnut features, where remission death is the target variable. AUC: area under the curve.

5.5 Discussion

In summary, this chapter considers the ability of several machine learning algorithms to accurately classify patients based on outcome. The algorithms were optimised according to hyperparameter tuning, whilst class weights and resampling methods were explored to handle the issues surrounding imbalanced data. Boosting and bagging techniques were also employed as "stronger learners" than the decision tree with the goal of reducing variance, bias and overfitting. Balanced weighting and leave-out-one cross validation were applied to the boosting algorithm to resolve imbalanced data issues whilst a larger weighting of the minority class was investigated with the bagging algorithm. A summary of the techniques and their success is presented in Table 50.

	Original data	Class weights	Under-	Over-	Under- and over-	LOO CV
Decision tree						
4 class outcome					In resampled data	
Rel/ref					In resampled data	
Remission death					In resampled data	
Random forest						
4 class outcome						
Rel/ref						
Remission death						
XGBoost						
4 class outcome						
Rel/ref						
Remission death						

Table 50. A grid summarising the machine learning algorithms, target variables, pruning techniques, and imbalanced data solutions employed for the Chestnut tree in this chapter, coloured by success. Only the Decision tree elements of the table were employed for the Oak and Elm trees. Green: successful, red: unsuccessful, grey: not employed. Techniques were considered successful if F1-scores of ≥ 80% were achieved in each class. LOO CV: leave-one-out cross-validation, rel/ref: relapse/ refractory disease, under: undersampled, over: oversampled.

The only decision trees able to accurately classify patients were those developed on resampled data. The Oak and Elm decision trees which used treatment elements as the features were largely unsuccessful even when applied to the resampled data. The Chestnut tree, which used total drug dosages and trial as the features, largely had the best performance with all three target variables in both genetic subgroups. The only successful Elm tree was that which classified *ETV6*::*RUNX1* patients by relapse/ refractory disease status, where undersampling had been performed to the data. However, even in this instance, the Chestnut tree had a better performance where under- and over-sampling had been implemented. The Oak tree was never successful.

It became clear that the good performance of the decision trees created with the resampled data did not translate to successful classification of patients in the original data, as in order to accurately classify a large proportion of the minority class, the algorithm also misclassified most of the majority class. The reason for this became clear when the data were visualised

using t-SNE plots with the total drug dosages and trial as features, as the clusters were not distinct based on outcome in the original data, likely causing difficulty in identifying splits within these data. Conversely, the resampled data showed largely distinct clusters between the relapse/ refractory disease cases and non-rel/ref cases which don't exist in the real data, thus illuminating the cause of the decision tree algorithm's superior performance in the resampled data. From these findings, it is evident that resampling data was not a solution to imbalanced data for this project. This issue is well reported within the literature, with studies finding that complex data, and data with lots of overlap of samples, is an issue for machine learning algorithms as resampling can cause overgeneralisation; thus models generated on resampled data can be poorly calibrated, both of which leads to a poor classification performance (Welvaars *et al.*, 2023; Kim and Jung, 2023; Sasada *et al.*, 2020). Despite this, resampling is still one of the most popular solutions to imbalanced data, with many different methods proposed, as few alternative solutions are available.

Bagging was utilised with the goal of reducing overfitting and increasing the prediction strength of the algorithm. Resampling techniques were not explored due to their proven inability to accurately represent the true data. The random forest algorithm did not improve upon the decision tree, once again misclassifying the minority group for each target variable in both *ETV6*::*RUNX1* and high hyperdiploidy populations. Balanced weight classes were employed to ensure the algorithm assigned equal importance to correctly classifying each group. This generally resulted in a marginal improvement in accuracy with regards to the minority class, but caused many of the majority class to be misclassified. Leave-one-out cross-validation was attempted as a method to give the algorithm as many true events as possible when training, with the aim of improving classification accuracy. When assessed in both the training and test data, the algorithms were once again unsuccessful; largely misclassifying the minority class. Thus, the classification performance of the random forest algorithm was poor overall and was not able to identify features which split patients based on outcome.

The final algorithm utilised was a boosting algorithm, which builds a weak learner and penalises incorrect classification. This algorithm performed similarly to the decision tree and random forest algorithms, by prioritising the classification of the majority class, resulting in a high accuracy overall, but misclassifying the minority class. A weighted XGBoost algorithm was employed which instructed the algorithm to weight the minority samples by 1000, thus

incentivising the model to prioritise these cases in its classification. This was unsuccessful in both *ETV6*::*RUNX1* and high hyperdiploidy datasets. Therefore, the XGBoost algorithm was also ineffective at achieving the aim of identifying features which could be utilised to predict the outcome of good risk patients.

Whilst the decision tree algorithm was successful in the resampled data, no machine learning algorithm utilised in this project was able to accurately classify patients based on outcome with the features used in the models in the original dataset. There are several potential causes for this. One possibility is that the algorithms utilised in the project weren't the most appropriate for the task. These algorithms were employed as they are often utilised for this type of task with positive results, and are also easily explainable – an important consideration when choosing machine learning algorithms as predictions made based on uninterpretable results is inappropriate in medical research. However, there are many machine learning algorithms that have been proposed for classification tasks within the literature, and further investigation into these may be necessary to resolve this problem (Sen, Hajra and Ghosh, 2020; Osisanwo *et al.*, 2017; Singh, Thakur and Sharma, 2016; Suyal and Goyal, 2022; Tan, 2021; Iqbal *et al.*, 2022). The algorithms utilised in this project were selected as they are fully explainable, meaning these algorithms show the steps made to classify patients. This is vital if these models are to be used in a real-world clinical setting. However, it is possible that this requirement of explainability has jeopardised the potential accuracy of these features to classify patients. Thus, alternative explainable or partially explainable algorithms may be more effective for this task whilst still maintaining their utility in a real-world setting. Another possible cause is that the data imbalance is too great such that no algorithm could classify the patients with the event of interest. Alternative resampling methods, weighting techniques, or decision thresholds could be explored both alone and in conjunction with different machine learning algorithms to resolve this issue (Esposito *et al.*, 2021; Tanha *et al.*, 2020; Puri and Kumar Gupta, 2022; Hasib *et al.*, 2020; Rawat and Mishra, 2024; Nakatsu, 2021). Alternatively, more data could be collected so that the algorithms have more cases to "learn" from and improve their classification performance. Another reason may be that there is too much homogeneity in the data, i.e. patients across groups have the same values in the features used by the model to assign classes, an issue that has been identified previously in the literature (Vuttipittayamongkol, Elyan and Petrovski, 2021). Finally, it is also possible, though unlikely,

that no optimal treatment elements/ drug dosages exist within the trials considered in this study. However, findings from chapter 3 and 4, as well as findings from previous studies dispute this notion (Enshaei *et al.*, 2021; Bartram, Veys and Vora, 2020; Samarasinghe *et al.*, 2021). A much more likely conclusion is that too few suitable features were available for consideration in this project.

Since a machine learning algorithm which accurately predicts outcome of good risks genetics patients could not be produced for this project, further work is required to develop an algorithm which can be used to identify optimal treatment elements on the four historic clinical trials. Furthermore, an alternative method to validate the findings of the relative dose intensity score is required.

Chapter 6. Discussion

6.1 Need for the study

Acute lymphoblastic leukaemia is the most common form of childhood cancer with good risk genetic abnormalities accounting for approximately 50% of B-cell precursor cases (Lustosa de Sousa *et al.*, 2015; Forestier *et al.*, 2008; Woodward *et al.*, 2023). Risk stratification and the development of intensified treatment protocols has dramatically improved the survival of these patients, resulting in cure rates >90% on contemporary protocols (Malard and Mohty, 2020; Inaba and Mullighan, 2020). This success has resulted in a growing number of survivors of childhood ALL (Andrés-Jensen *et al.*, 2020). However, patients often experience acute toxicities causing ~40% of deaths during treatment, whilst survivors suffer long-term late effects as a result of this treatment, with them being expected to have at least one chronic condition by the age of 40 (Al-Mahayri, AlAhmad and Ali, 2021; Andrés-Jensen *et al.*, 2020). Therapy-related complications are widespread and can affect many organ systems including psychosocial problems, neurocognitive deficits, cardiovascular issues, musculoskeletal problems, auditory and ocular impairments, reproductive issues, and subsequent malignant neoplasms (Landier *et al.*, 2018).

Treatment of patients has been modified over the past few decades in order to lessen the burden; omitting or reducing treatment elements known to cause particular late effects (Dixon *et al.*, 2020). Cranial radiation therapy is an example of this, as it has been removed from CNS treatment for the most part due to its harmful neurocognitive effects (Al-Mahayri, AlAhmad and Ali, 2021). Whilst these changes to therapy have resulted in fewer instances of the late effects that were seen in patients treated before 1990 - specifically reproductive, neurological, or gastrointestinal effects and CRT induced hypothalamic dysfunction; survivors treated with conventional therapy still experience impaired glucose metabolism and obesity as well as musculoskeletal effects, possibly due to more intensive dexamethasone and asparaginase doses (Inaba and Mullighan, 2020).

Furthermore, several treatment-related risk factors surrounding common chemotherapy drugs have been reported. These include asparaginase toxicities such as pancreatitis, thrombosis, and hepatotoxicity, as well as adverse effects as a result of dexamethasone treatment encompassing neuropsychological and metabolic issues (Gupta *et al.*, 2020; Warris *et al.*, 2016). Moreover, the neurotoxic effect of vincristine, cardiovascular disease prevalence of patients treated with anthracycline, bone toxicity caused by corticosteroids and

methotrexate, as well as an association between anthracycline and secondary malignancies including breast cancer in female survivors and acute myeloid leukaemia have also been reported (Al-Mahayri, AlAhmad and Ali, 2021). As such, studies have focused increasingly on identifying measures to reduce toxicity and long-term sequelae in patients such as introducing other agents to mitigate adverse events, individualising drug dosages through therapeutic drug monitoring, or identifying patients with an increased risk to toxic effects and reducing their exposure to certain medications (Warris *et al.*, 2016; Kloos *et al.*, 2020; Pui and Evans William, 2006). Additionally, it has been shown that risk-stratified therapy, which aims to improve survival by increasing the intensity of therapy for high risk patients whilst reducing therapy for standard risk patients, has also succeeded in reducing late morbidity and mortality for standard risk patients (Dixon *et al.*, 2020).

Recent studies have shown that patients with low risk features such as good risk genetics and low levels of MRD at the end of induction could be eligible for further de-escalation of therapy (Moorman *et al.*, 2022a; Pedrosa *et al.*, 2020; Sidhom *et al.*, 2021; Li *et al.*, 2021; Goodwin, 2023). The rationale of this study is that optimal treatment elements/ therapy intensities already exist for these patients within historic clinical trial protocols due to high and stable cure rates in the good risk genetics subgroups over the previous consecutive clinical trials. Previous studies have looked to identify optimal therapy in Europe using a vertical approach of contemporary clinical trials, however no-one has used a horizontal approach as employed in this study (Østergaard *et al.*, 2024).

6.2 Summary of findings

It is vital to find optimal treatment pathways for cure of ALL whilst minimising toxicities as much as possible to improve long-term late effects of survivors. This is especially important in ALL as it is predominantly a childhood disease, thus there is a lifelong risk of developing and living with these sequelae. This study aimed to identify treatment elements that are optimal for good risk genetic patients to ensure they are given only the minimal dosages necessary for cure, whilst reducing unnecessary toxicities.

A dataset comprising of patient information from a cohort of paediatric ALL patients treated on four consecutive clinical trials was assembled through collation of data available from LRCG data sources. Key treatment elements trial, regimen, and delayed intensification were

assessed for their impact on outcome of good risk genetics patients. These investigations revealed that outcome of *ETV6*::*RUNX1* patients were comparable from UKALL97 onwards whilst high hyperdiploidy patients had stable cure rates from UKALL97/99; suggesting that there exists an optimal pathway within these trials and that the treatment de-escalation received in the latter trials did not adversely affect outcomes of good risk genetics patients. Furthermore, the increase in good risk genetics patients treated on regimen C across the latter three trials was found to be unnecessary, due to stable survival rates within each regimen across the trials. Analysis of survival by number of delayed intensifications showed that by UKALL97, the addition of the third intensification block was of detriment to patients with good risk genetics, and that the reduction to one delayed intensification on UKALL2003 and UKALL2011 did not significantly affect survival when compared to patients receiving two delayed intensifications on UKALL2003. This is evidence that the de-intensification of treatment protocols for good risk genetics patients was not detrimental to survival (Chapter 3).

Drug dosages for these patients were calculated from trial protocols and appended to the dataset. A novel dose intensity score, based on the work of Hryniuk *et al.*, was calculated which successfully classifies patients by dose intensity. Survival analysis of quartiles of the relative dose intensity score showed that the very low intensity treatment arms resulted in optimal survival of both *ETV6*::*RUNX1* and high hyperdiploidy patients whilst patients treated with high dose methotrexate had inferior survival. However, the high dose intensity score group was enriched with UKALLXI92 patients, shown to have the poorest survival rates of all patients in this study. Thus, further analysis is required to ensure this result is independent of the effects of other treatment elements. Furthermore, it was found that methods to calculate a dose intensity score proposed by Hryniuk and Allgoewer were not efficacious in this cohort (Chapter 4).

Machine learning algorithms were utilised as an alternative method to identify optimal dosages and treatment elements for cure of patients; seeking to validate the findings of the previous analyses in this study. Three machine learning algorithms were explored with various pruning and resampling techniques investigated. These algorithms were ultimately unsuccessful, likely due to heavily imbalanced data that was homogenous in its features. This was supported by a dimension reduction visualisation technique (t-SNE) which showed that

the patients with an event of interest clustered with the non-event patients when these features were used. Further work to resolve the data imbalance and identify a more robust algorithm is required (Chapter 5).

In summary, this study provides further evidence that low-risk patients, namely those with an *ETV6*::*RUNX1* translocation or high hyperdiploidy, are suitable for treatment de-escalation through analysis of treatment elements on historic consecutive UKALL clinical trials, as well as through analysis of a novel dose intensity score. Furthermore, this study establishes the use of machine learning to identify optimal treatment elements for patients with ALL. The impact of these findings is explored in Section 6.3.

6.3 Relevance of findings and context within the field

Subgroups of patients with cure rates >90% on contemporary trials have been identified highlighting the opportunity for treatment de-escalation in these populations. However, the minimal amount of therapy needed for cure of ALL, as well as which elements of treatment are essential to cure, are yet to be established.

Analysis of treatment elements of four clinical trials identified that one delayed intensification is sufficient for treatment of good risk genetic ALL patients when compared against 2 delayed intensifications (*ETV6*::*RUNX1* 5-year OS rates: 99% vs 96%, high hyperdiploidy 5 year OS rates: 99% and 94%) and that additional intensifications were detrimental to survival. Furthermore, it was found that allocating patients with good risk genetics to regimen C due to MRD status at day 28 was of no benefit, with comparable survival rates of patients on each regimen across the three trials [Tables 13-16]. This confirms the findings of Østergaard *et al.* who found that *ETV6*::*RUNX1* patients had excellent survival rates irrespective of MRD stratification (Østergaard *et al.*, 2024). Of note is that patients treated on UKALL2011 did have slightly inferior EFS rates than those treated on UKALL2003 in both good risk genetic subgroups, however this difference was not significant, nor did it impair cure rates. Importantly, the treatment backbone in UKALL2011 was less intensive than UKALL2003, thus the observation that survival was equivalent provides further evidence that patients with good risk genetics can, on average, be cured with less chemotherapy. Thus, one could argue that the benefit that the overall de-intensification provides for all good risk genetics patients outweighs the additional events, resulting in relapse salvaging treatment, seen in a minority

of cases. These findings confer with other study groups whom have determined that standard risk patients do not benefit from more intensive treatment (Hunger *et al.*, 2013; Schore *et al.*, 2023; Maloney *et al.*, 2019; Sidhom *et al.*, 2021).

Investigation of the relative dose intensity score supported these findings with the very low intensity risk group (comprised of every patient receiving one delayed intensification, as well as the subgroup of patients receiving two delayed intensifications and intrathecal methotrexate on UKALLXI92) having superior survival rates in both overall and event-free endpoints [Figures 45 - 48]. These findings suggest that patients were likely over-treated on the other treatment protocols, and that the treatment reduction seen on UKALL2011 was suitable for good risk genetic subgroups. This once again supports the findings of Østergaard *et al.* who determined that treatment de-escalation should be considered for the majority of *ETV6*::*RUNX1* patients due to the similar outcomes seen on several contemporary clinical trials irrespective of treatment intensity (Østergaard *et al.*, 2024). Similarly, that study concluded that certain treatment elements, in particular glucocorticoid and vincristine pulses during maintenance as well as the use of HD-MTX, may not be necessary for treatment of *ETV6*::*RUNX1* patients, and further proposed the reduction of anthracyclines. The results of this study also suggest that HD-MTX is unnecessary in the treatment of patients with good risk genetics, as these patients were assigned to the high intensity group, which had the worst outcomes in both endpoints [Figures 45 and 46]. Several other studies have also investigated the effect of vincristine-steroid pulses during maintenance, concluding that they offer no additional benefit to overall survival (Guolla *et al.*, 2023; Childhood Acute Lymphoblastic Leukaemia Collaborative, 2010; Hinze *et al.*, 2017). Moreover, studies have found that anthracycline dosage can be reduced in patients with low levels of MRD and *ETV6*::*RUNX1* patients (Pieters *et al.*, 2016; Schrappe *et al.*, 2017; Pieters *et al.*, 2023). Further analysis is required to determine the effect of dose reduction of individual drugs in this study cohort.

Importantly, this study presents a novel approach to calculating the dose intensity of patients. This can be utilised to offer further insight into optimal dosages of patients, as well as the feasibility of de-escalation in multiple populations; an important step in the direction of personalised therapy. Furthermore, the development of a dataframe of daily drug dosages for patients treated on four consecutive UKALL trials allows for avenues of research previously unavailable to this cohort. Finally, this work provides a foundation for the exploration into the

use of machine learning algorithms for identifying ALL patients eligible for treatment de-escalation; substantiating this approach to analysis in cancer research. This type of study could complement ongoing research in the field which, in addition to other applications, works to detect haematological disorders, classify patients into subtypes, predict the likelihood of relapse in newly diagnosed patients, and identify significant clinical and phenotypic risk factors (Das *et al.*, 2022; Mahmood *et al.*, 2020; Rehman *et al.*, 2018; Das, Pradhan and Meher, 2021; Pan *et al.*, 2017).

6.4 Study strengths and limitations

This study assembled a cohort of 9163 patients (including all genetic subtypes) with detailed treatment information and appended these data with daily drug dosage values, producing a resource for future analyses. The size of this cohort ensures strong statistical power and allows for further sub-classification of patients for subgroup specific analyses and comparison. A further benefit of this study is the development of a novel dose intensity score to aid in the classification of patients as well as for use in determining the minimal intensity of treatment required for cure. As dosages were employed directly into the model in the units with which they were prescribed, this allows for straightforward applicability to other datasets and studies. Finally, this study assessed the possibility that a treatment pathway on historic clinical trials may be optimal for cure of good risk genetic patients through its horizontal approach to the analysis, which is a novel approach to this type of research.

There are several limitations to this study. Firstly, information regarding individual patients' administered drug dosages were unavailable and were instead calculated from the protocols with the assumption that no major deviations from the treatment plan occurred. However, this is reasonable, as patients who experienced major deviations from the prescribed dose were often taken off trial, a factor that was accounted for in this study. A related limitation is that values were assigned for drugs which were prescribed with an escalating dose to toxicity, as this information was also not collected. Although these values were randomly assigned, a logical range was determined and a Normal distribution was used to assign the values, centred on the rationale that most patients will tolerate some escalation, but few will achieve the maximum possible dose, thus an approximately Normal distribution is to be expected. Furthermore, certain drugs appeared to have a disproportionate effect on the dose intensity scores due to the large variation in doses administered – in particular high dose methotrexate.

Thus, the relative dose intensity score will likely need modification to account for this factor before being taken forward.

A further limitation to this study was the lack of pharmacokinetic, pharmacodynamic and pharmacogenomic data available. Thus, individual patient response to treatment was not able to be assessed, nor was the interactions between drugs at differing intensities. Therefore, it was assumed in this study that patient response was equivalent across the cohort, implying that dose intensity was stable amongst patients, not affected by drug-drug interactions, and bore no effect on toxicities; a known falsehood with several studies describing these effects within ALL treatment (Yang *et al.*, 2008; Groninger *et al.*, 2002; Csordas *et al.*, 2013; Kawedia *et al.*, 2011).

Finally, one limitation to the machine learning element of this study was directly related to the imbalanced nature of the data. This imbalance stems from the fact that good risk patients have few events, resulting in an insignificant number of positive instances in the data. The extreme data imbalance seen within this cohort is classed as severely imbalanced, defined by majority-to-minority class ratios between 100:1 and 10,000:1 (Hasanin *et al.*, 2019). This often results in machine learning algorithms having an inability to distinguish between majority and minority class samples due to a lack of examples to draw from to learn characteristic features of the minority class. Furthermore, with such a large imbalance, the algorithm can completely misclassify the minority class with little detriment to the overall accuracy score of the model, as only ~1% of cases are being incorrectly classified. As such, no algorithm was developed that could accurately classify patients by outcome using drug dosages, thus the findings of the relative dose intensity score could not be validated in this study.

6.5 Future work

The effect of different treatment elements on outcome were analysed separately in this project to assess the optimal number of delayed intensification required for treatment of good risk genetics patients, as well as to assess the benefit of assigning patients to regimens based on MRD status at the end of induction. However, further analysis should be performed to assess the combined effect of these elements through multivariate analysis, accounting for any possible interactions. This is essential as differences in outcome seen by these elements

may not be as a result of this factor independently, and instead may be influenced by several factors simultaneously, as no treatment element was used singularly.

During this study, drug dosages for patients treated on four paediatric UKALL clinical trials were calculated using trial protocol information and appended to existing data. A recommended additional step would be to assess the validity of the assumptions made in this study by retrieving data from electronic health records held by the NHS regarding the actual doses administered to patients. A feasible approach to this would be to secure this data for a representative subset of patients and determine if their estimated doses align with their actual doses. If there is no significant difference between the values for this representative cohort, then it could be assumed that this is the case for all patients in this study. A long-term goal would be to acquire these data for all patients in the study and update the dataframe with the accurate values and recalculate the dose intensity score to reflect this. This would further the advancement towards individualised therapy with each patient having an intensity score calculated based on the amount of therapy tolerated.

Although not possible in this study due to the lack of data, validation of the dose intensity score should be performed in a dataset for which pharmaco-kinetic, -dynamic, and –genomic data are available. Therefore, one could determine if the dose intensity score accurately appropriates the true intensity received by patients when individual drug metabolisation is accounted for in the analysis. Additionally, if the dose intensity score cannot be validated within this cohort, these data could be used as variables to build an alternative dose intensity score. This would improve upon the accuracy of the dose intensity score, helping to assign therapy more precisely as well as allow for each patient's personal response to therapy to be a factor in their treatment.

Lastly, a preferable approach to resolve the issue of imbalanced data for machine learning would be to obtain more data of positive cases from other studies. This would allow the algorithms to have more cases to learn from and improve classification of these patients of interest. Should real-world data be unavailable, one could consider the use of synthetically generated data as was exemplified by Eckardt *et al.* in acute myeloid leukaemia (Eckardt *et al.*, 2024). An alternative avenue would be to include more features for the model to consider, or use samples with a wider array of values in the original features in order to give the

algorithm more heterogeneity in the feature space; an important component of model building in machine learning (Xiao and Wang, 2019; Woodward *et al.*, 2022; Sahoo *et al.*, 2022). This results in data for which there are differences in the features between the classes, allowing the algorithm to split the classes more effectively.

Should additional data be unavailable, other machine learning algorithms or resampling/ weighting techniques could be explored. There are a wealth of classification algorithms each with different approaches and strengths, making them particularly suited to certain tasks (Singh, Thakur and Sharma, 2016; Suyal and Goyal, 2022; Sen, Hajra and Ghosh, 2020). It was only possible to assess the utility of three algorithms during this project, however many others could be applied to this data such as artificial neural networks, logistic regression, support vector machine (SVM), naïve bayes, and K-nearest neighbour (KNN) models.

Furthermore, as imbalanced data is a major issue in real-world applications of ML algorithms, various methods have been proposed to solve this issue which can be classified into four groups: (1) data pre-processing or resampling methods, (2) algorithm level methods, (3) ensemble methods, and (4) cost-sensitive methods. Methods 1, 3, and 4 were explored within this project in the form of under- and over-sampling, ensemble methods, and class weighting respectively. However, algorithm level methods have yet to be explored. Therefore, approaches to modify the underlying learner to reduce bias towards the majority group could be explored (Johnson and Khoshgoftaar, 2019). Alternative techniques include adjusting the decision threshold to an optimal threshold according to a balanced accuracy metric (such as the F1-score), utilising a one-class classifier which learns how to identify only the minority class, thus giving the model the ability to separate them from outliers (i.e. the majority class), or algorithms developed specifically to handle imbalanced data (Esposito *et al.*, 2021; Fernández *et al.*, 2018a; Puri and Kumar Gupta, 2022).

6.6 Final summary

In summary, this study has successfully assembled a dataframe of both drug dosage and patient-level information including demographic, genetic, treatment, and outcome data for patients on four consecutive clinical trials. Furthermore, a novel dose intensity score has been developed which can be utilised in further studies to identify optimal doses for cure of ALL patients. Survival analysis of treatment elements and dose intensity has confirmed that

treatment de-escalation is possible for patients with good risk genetics and the outcome of these patients on both UKALL2003 and UKALL2011 supports the use of one delayed intensification and lower intensity regimens in these subgroups. Additionally, this thesis provides a basis for machine learning to be used alongside traditional statistical methods in determining optimal treatment elements and intensity. These findings advance efforts towards personalised therapy with the goal of reducing toxicity and long-term late effects in children treated for acute lymphoblastic leukaemia.

Chapter 7. References

Abajian, A., Murali, N., Savic, L. J., Laage-Gaupp, F. M., Nezami, N., Duncan, J. S., Schlachter, T., Lin, M., Geschwind, J. F. and Chapiro, J. (2018) 'Predicting Treatment Response to Intra-arterial Therapies for Hepatocellular Carcinoma with the Use of Supervised Machine Learning-An Artificial Intelligence Concept', *J Vasc Interv Radiol,* 29(6), pp. 850-857.e1.

Ahmad, G. N., Fatima, H., Ullah, S., Saidi, A. S. and Imdadullah (2022) 'Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV', *IEEE Access,* 10, pp. 80151-80173.

Al-Mahayri, Z. N., AlAhmad, M. M. and Ali, B. R. (2021) 'Long-term effects of pediatric acute lymphoblastic leukemia chemotherapy: can recent findings inform old strategies?', *Frontiers in Oncology,* 11, pp. 710163.

Allgoewer, A., Schmid, M., Radermacher, P., Asfar, P. and Mayer, B. (2018) 'Area under the curve-derived measures characterizing longitudinal patient responses for given thresholds', *Epidemiology, Biostatistics, and Public Health,* 15(4).

Alpar, D., Wren, D., Ermini, L., Mansur, M. B., van Delft, F. W., Bateman, C. M., Titley, I., Kearney, L., Szczepanski, T., Gonzalez, D., Ford, A. M., Potter, N. E. and Greaves, M. (2015) 'Clonal origins of ETV6-RUNX1+ acute lymphoblastic leukemia: studies in monozygotic twins', *Leukemia,* 29(4), pp. 839-846.

Alvarez, Y., Caballín, M. R., Gaitán, S., Pérez, A., Bastida, P., Ortega, J. J., Cervera, J., Verdeguer, A., Tasso, M., Aventín, A., Badell, I., Guitart, M., Melo, M., Granada, I., Javier, G., Dastugue, N., Robert, A. and Coll, M. D. (2007) 'Presenting features of 201 children with acute lymphoblastic leukemia: Comparison according to presence or absence of ETV6/RUNX1 rearrangement', *Cancer Genetics and Cytogenetics,* 177(2), pp. 161-163.

Ampatzidou, M., Papadhimitriou, S. I., Paterakis, G., Pavlidis, D., Tsitsikas, K., Kostopoulos, I. V., Papadakis, V., Vassilopoulos, G. and Polychronopoulou, S. (2018) 'ETV6/RUNX1-positive childhood acute lymphoblastic leukemia (ALL): The spectrum of clonal heterogeneity and its impact on prognosis', *Cancer Genetics,* 224-225, pp. 1-11.

Amylon, M. D., Shuster, J., Pullen, J., Berard, C., Link, M. P., Wharam, M., Katz, J., Yu, A., Laver, J., Ravindranath, Y., Kurtzberg, J., Desai, S., Camitta, B. and Murphy, S. B. (1999) 'Intensive high-dose asparaginase consolidation improves survival for pediatric patients with T cell acute lymphoblastic leukemia and advanced stage lymphoblastic lymphoma: a Pediatric Oncology Group study', *Leukemia,* 13(3), pp. 335-342.

Andrés-Jensen, L., Larsen, H. B., Johansen, C., Frandsen, T. L., Schmiegelow, K. and Wahlberg, A. (2020) 'Everyday life challenges among adolescent and young adult survivors of childhood acute lymphoblastic leukemia: An in-depth qualitative study', *Psycho-Oncology,* 29(10), pp. 1630-1637.

Aplenc, R., Glatfelter, W., Han, P., Rappaport, E., La, M., Cnaan, A., Blackwood, M. A., Lange, B. and Rebbeck, T. (2003) 'CYP3A genotypes and treatment response in paediatric acute lymphoblastic leukaemia', *British Journal of Haematology,* 122(2), pp. 240-244.

Aricò, M., Valsecchi, M. G., Rizzari, C., Barisone, E., Biondi, A., Casale, F., Locatelli, F., Lo Nigro, L., Luciani, M., Messina, C., Micalizzi, C., Parasole, R., Pession, A., Santoro, N.,

Testi, A. M., Silvestri, D., Basso, G., Masera, G. and Conter, V. (2008) 'Long-term results of the AIEOP-ALL-95 Trial for Childhood Acute Lymphoblastic Leukemia: insight on the prognostic value of DNA index in the framework of Berlin-Frankfurt-Muenster based chemotherapy', *J Clin Oncol,* 26(2), pp. 283-9.

Armand, P., Kim, H. T., Logan, B. R., Wang, Z., Alyea, E. P., Kalaycio, M. E., Maziarz, R. T., Antin, J. H., Soiffer, R. J., Weisdorf, D. J., Rizzo, J. D., Horowitz, M. M. and Saber, W. (2014) 'Validation and refinement of the Disease Risk Index for allogeneic stem cell transplantation', *Blood,* 123(23), pp. 3664-3671.

Arora, S., Hu, W. and Kothari, P. K. 'An Analysis of the t-SNE Algorithm for Data Visualization', *Proceedings of the 31st Conference On Learning Theory*, Proceedings of Machine Learning Research: PMLR, 1455--1462.

Austin, M. and Patel, B. (2023) 'The Acute Leukaemias', *ABC of Clinical Haematology*, pp. 31.

Bailey, L. C., Lange, B. J., Rheingold, S. R. and Bunin, N. J. (2008) 'Bone-marrow relapse in paediatric acute lymphoblastic leukaemia', *The Lancet Oncology,* 9(9), pp. 873-883.

Bain, B. J. (2017) *Leukaemia diagnosis.* John Wiley & Sons.

Banegas-Luna, A. J., Peña-García, J., Iftene, A., Guadagni, F., Ferroni, P., Scarpato, N., Zanzotto, F. M., Bueno-Crespo, A. and Pérez-Sánchez, H. (2021) 'Towards the Interpretability of Machine Learning Predictions for Medical Applications Targeting Personalised Therapies: A Cancer Case Survey', *International Journal of Molecular Sciences*, 22(9). DOI: 10.3390/ijms22094394.

Barnes, E. (2008) 'Cancer coverage: the public face of childhood leukaemia in 1960s Britain', *Endeavour,* 32(1), pp. 10-15.

Bartram, J., Veys, P. and Vora, A. (2020) 'Improvements in outcome of childhood acute lymphoblastic leukaemia (ALL) in the UK – a success story of modern medicine through successive UKALL trials and international collaboration', *British Journal of Haematology,* 191(4), pp. 562-567.

Bayat Mokhtari, R., Homayouni, T. S., Baluch, N., Morgatskaya, E., Kumar, S., Das, B. and Yeger, H. (2017) 'Combination therapy in combating cancer', *Oncotarget,* 8(23), pp. 38022-38043.

Bennett, K. (1995) 'Global Tree Optimization: A Non-greedy Decision Tree Algorithm', *Computing Sciences and Statistics,* 26.

Berry, D. A., Zhou, S., Higley, H., Mukundan, L., Fu, S., Reaman, G. H., Wood, B. L., Kelloff, G. J., Jessup, J. M. and Radich, J. P. (2017) 'Association of Minimal Residual Disease With Clinical Outcome in Pediatric and Adult Acute Lymphoblastic Leukemia: A Meta-analysis', *JAMA Oncology,* 3(7), pp. e170580-e170580.

Bhakta, N., Liu, Q., Ness, K. K., Baassiri, M., Eissa, H., Yeo, F., Chemaitilly, W., Ehrhardt, M. J., Bass, J., Bishop, M. W., Shelton, K., Lu, L., Huang, S., Li, Z., Caron, E., Lanctot, J., Howell, C., Folse, T., Joshi, V., Green, D. M., Mulrooney, D. A., Armstrong, G. T., Krull, K. R., Brinkman, T. M., Khan, R. B., Srivastava, D. K., Hudson, M. M., Yasui, Y. and Robison, L. L. (2017) 'The cumulative burden of surviving childhood cancer: an initial report from the St Jude Lifetime Cohort Study (SJLIFE)', *The Lancet,* 390(10112), pp. 2569-2582.

Bhojwani, D., Pei, D., Sandlund, J. T., Jeha, S., Ribeiro, R. C., Rubnitz, J. E., Raimondi, S. C., Shurtleff, S., Onciu, M., Cheng, C., Coustan-Smith, E., Bowman, W. P., Howard, S. C., Metzger, M. L., Inaba, H., Leung, W., Evans, W. E., Campana, D., Relling, M. V. and Pui, C. H. (2012) 'ETV6-RUNX1-positive childhood acute lymphoblastic leukemia: improved outcome with contemporary therapy', *Leukemia,* 26(2), pp. 265-270.

Biau, G. and Scornet, E. (2016) 'A random forest guided tour', *TEST,* 25(2), pp. 197-227.

Bleyer, W. A., Nelson, J. A. and Kamen, B. A. (1997) 'Accumulation of Methotrexate in Systemic Tissues after Intrathecal Administration', *Journal of Pediatric Hematology/Oncology,* 19(6).

Blockeel, H., Struyf, J., Brodley, E. and Danyluk, A. (2003) 'Journal of Machine Learning Research 3 (2002) 621-650 Submitted 12/01; Published 12/02 Efficient Algorithms for Decision Tree Cross-validation'.

Borowitz, M. J., Devidas, M., Hunger, S. P., Bowman, W. P., Carroll, A. J., Carroll, W. L., Linda, S., Martin, P. L., Pullen, D. J., Viswanatha, D., Willman, C. L., Winick, N., Camitta, B. M. and for the Children's Oncology, G. (2008) 'Clinical significance of minimal residual disease in childhood acute lymphoblastic leukemia and its relationship to other prognostic factors: a Children's Oncology Group study', *Blood,* 111(12), pp. 5477-5485.

Bostrom, B. C., Erdmann, G. R. and Kamen, B. A. (2003) 'Systemic Methotrexate Exposure Is Greater After Intrathecal Than After Oral Administration', *Journal of Pediatric Hematology/Oncology,* 25(2).

Boulesteix, A.-L., Janitza, S., Kruppa, J. and König, I. R. (2012) 'Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics', *WIREs Data Mining and Knowledge Discovery,* 2(6), pp. 493-507.

Braun, T. P., Eide, C. A. and Druker, B. J. (2020) 'Response and Resistance to BCR-ABL1-Targeted Therapies', *Cancer Cell,* 37(4), pp. 530-542.

Breiman, L. (1996) 'Bagging predictors', *Machine learning,* 24, pp. 123-140.

Brody, T. (2012) 'Chapter 9 - Biostatistics', in Brody, T. (ed.) *Clinical Trials*. Boston: Academic Press, pp. 165-190.

Brown, P. A., Ji, L., Xu, X., Devidas, M., Hogan, L. E., Borowitz, M. J., Raetz, E. A., Zugmaier, G., Sharon, E., Bernhardt, M. B., Terezakis, S. A., Gore, L., Whitlock, J. A., Pulsipher, M. A., Hunger, S. P. and Loh, M. L. (2021) 'Effect of Postreinduction Therapy Consolidation With Blinatumomab vs Chemotherapy on Disease-Free Survival in Children, Adolescents, and Young Adults With First Relapse of B-Cell Acute Lymphoblastic Leukemia: A Randomized Clinical Trial', *JAMA,* 325(9), pp. 833-842.

Buchmann, S., Schrappe, M., Baruchel, A., Biondi, A., Borowitz, M., Campbell, M., Cario, G., Cazzaniga, G., Escherich, G., Harrison, C. J., Heyman, M., Hunger, S. P., Kiss, C., Liu, H.-C., Locatelli, F., Loh, M. L., Manabe, A., Mann, G., Pieters, R., Pui, C.-H., Rives, S., Schmiegelow, K., Silverman, L. B., Stary, J., Vora, A., Brown, P. and on behalf of the Ponte-di-Legno, C. (2022) 'Remission, treatment failure, and relapse in pediatric ALL: an international consensus of the Ponte-di-Legno Consortium', *Blood,* 139(12), pp. 1785-1793.

Butler, R. W. and Haser, J. K. (2006) 'Neurocognitive effects of treatment for childhood cancer', *Mental Retardation and Developmental Disabilities Research Reviews,* 12(3), pp. 184-191.

Campana, D. (2010) 'Minimal Residual Disease in Acute Lymphoblastic Leukemia', *Hematology,* 2010(1), pp. 7-12.

Cancela, C. S., Murao M Fau - Viana, M. B., Viana Mb Fau - de Oliveira, B. M. and de Oliveira, B. M. (2012) 'Incidence and risk factors for central nervous system relapse in children and adolescents with acute lymphoblastic leukemia', *Revista Brasileira de Hematologia e Hemoterapia,* (1516-8484 (Print)).

Cancer Research UK (2024a) *Acute lymphoblastic leukaemia (ALL) incidence statistics*. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/leukaemia-all/incidence 2024).

Cancer Research UK (2024b) *Children's Cancers Incidence Statistics*. Available at: https://www.cancerresearchuk.org/health-professional/cancer-statistics/childrens-cancers/incidence#ref- (Accessed: 2024.

Ceppi, F., Langlois-Pelletier, C., Gagné, V., Rousseau, J., Ciolino, C., Lorenzo, S. D., Kevin, K. M., Cijov, D., Sallan, S. E., Silverman, L. B., Neuberg, D., Kutok, J. L., Sinnett, D., Laverdière, C. and Krajinovic, M. (2014) 'Polymorphisms of the Vincristine Pathway and Response to Treatment in Children with Childhood Acute Lymphoblastic Leukemia', *Pharmacogenomics,* 15(8), pp. 1105-1116.

Chang, T.-C., Chen, W., Qu, C., Cheng, Z., Hedges, D., Elsayed, A., Pounds, S. B., Shago, M., Rabin, K. R., Raetz, E. A., Devidas, M., Cheng, C., Angiolillo, A., Baviskar, P., Borowitz, M., Burke, M. J., Carroll, A., Carroll, W. L., Chen, I. M., Harvey, R., Heerema, N., Iacobucci, I., Wang, J. R., Jeha, S., Larsen, E., Mattano, L., Maloney, K., Pui, C.-H., Ramirez, N. C., Salzer, W., Willman, C., Winick, N., Wood, B., Hunger, S. P., Wu, G., Mullighan, C. G. and Loh, M. L. (2024) 'Genomic Determinants of Outcome in Acute Lymphoblastic Leukemia', *Journal of Clinical Oncology,* 0(0), pp. JCO.23.02238.

Charbuty, B. and Abdulazeez, A. (2021) 'Classification Based on Decision Tree Algorithm for Machine Learning', *Journal of Applied Science and Technology Trends,* 2(01), pp. 20 - 28.

Chawla, N., Bowyer, K., Hall, L. and Kegelmeyer, W. (2002) 'SMOTE: Synthetic Minority Over-sampling Technique', *J. Artif. Intell. Res. (JAIR),* 16, pp. 321-357.

Chen, J., Gale, R. P., Hu, Y., Yan, W., Wang, T. and Zhang, W. (2024) 'Measurable residual disease (MRD)-testing in haematological and solid cancers', *Leukemia,* 38(6), pp. 1202-1212.

Chennamaneni, R., Gundeti, S., Konatam, M. L., Bala, S., Kumar, A. and Srinivas, L. (2018) 'Impact of cytogenetics on outcomes in pediatric acute lymphoblastic leukemia', *South Asian Journal of Cancer,* (2278-330X (Print)).

Cheok, M. H. and Evans, W. E. (2006) 'Acute lymphoblastic leukaemia: a model for the pharmacogenomics of cancer therapy', *Nature Reviews Cancer,* 6(2), pp. 117-129.

Cheok, M. H., Pottier, N., Kager, L. and Evans, W. E. (2009) 'Pharmacogenetics in Acute Lymphoblastic Leukemia', *Seminars in Hematology,* 46(1), pp. 39-51.

Childhood Acute Lymphoblastic Leukaemia Collaborative, G. (2010) 'Systematic review of the addition of vincristine plus steroid pulses in maintenance treatment for childhood acute lymphoblastic leukaemia – an individual patient data meta-analysis involving 5659 children', *British Journal of Haematology,* 149(5), pp. 722-733.

Childhood Acute Lymphoblastic Leukaemia Collaborative Group (2009) 'Beneficial and harmful effects of anthracyclines in the treatment of childhood acute lymphoblastic leukaemia: a systematic review and meta-analysis', *British Journal of Haematology,* 145(3), pp. 376-388.

Chilton, L., Buck, G., Harrison, C. J., Ketterling, R. P., Rowe, J. M., Tallman, M. S., Goldstone, A. H., Fielding, A. K. and Moorman, A. V. (2014) 'High hyperdiploidy among adolescents and adults with acute lymphoblastic leukaemia (ALL): cytogenetic features, clinical characteristics and outcome', *Leukemia,* 28(7), pp. 1511-1518.

Chin Neoh, S., Srisukkham, W., Zhang, L., Todryk, S., Greystoke, B., Peng Lim, C., Alamgir Hossain, M. and Aslam, N. (2015) 'An Intelligent Decision Support System for Leukaemia Diagnosis using Microscopic Blood Images', *Scientific Reports,* 5(1), pp. 14938.

Clappier, E., Grardel, N., Bakkus, M., Rapion, J., De Moerloose, B., Kastner, P., Caye, A., Vivent, J., Costa, V., Ferster, A., Lutz, P., Mazingue, F., Millot, F., Plantaz, D., Plat, G., Plouvier, E., Poirée, M., Sirvent, N., Uyttebroeck, A., Yakouben, K., Girard, S., Dastugue, N., Suciu, S., Benoit, Y., Bertrand, Y., Cavé, H. and on behalf of the, E.-C. (2015) 'IKZF1 deletion is an independent prognostic marker in childhood B-cell precursor acute lymphoblastic leukemia, and distinguishes patients benefiting from pulses during maintenance therapy: results of the EORTC Children's Leukemia Group study 58951', *Leukemia,* 29(11), pp. 2154-2161.

Clark, T. G., Bradburn, M. J., Love, S. B. and Altman, D. G. (2003) 'Survival Analysis Part I: Basic concepts and first analyses', *British Journal of Cancer,* 89(2), pp. 232-238.

Conter, V., Bartram, C. R., Valsecchi, M. G., Schrauder, A., Panzer-Grümayer, R., Möricke, A., Aricò, M., Zimmermann, M., Mann, G., De Rossi, G., Stanulla, M., Locatelli, F., Basso, G., Niggli, F., Barisone, E., Henze, G., Ludwig, W.-D., Haas, O. A., Cazzaniga, G., Koehler, R., Silvestri, D., Bradtke, J., Parasole, R., Beier, R., van Dongen, J. J. M., Biondi, A. and Schrappe, M. (2010) 'Molecular response to treatment redefines all prognostic factors in children and adolescents with B-cell precursor acute lymphoblastic leukemia: results in 3184 patients of the AIEOP-BFM ALL 2000 study', *Blood,* 115(16), pp. 3206-3214.

Cooper, S. L. and Brown, P. A. (2014) 'Treatment of pediatric acute lymphoblastic leukemia', *Pediatric Clinics of North America,* (1557-8240 (Electronic)).

Coustan-Smith, E., Mullighan, C. G., Onciu, M., Behm, F. G., Raimondi, S. C., Pei, D., Cheng, C., Su, X., Rubnitz, J. E., Basso, G., Biondi, A., Pui, C.-H., Downing, J. R. and Campana, D. (2009) 'Early T-cell precursor leukaemia: a subtype of very high-risk acute lymphoblastic leukaemia', *The Lancet Oncology,* 10(2), pp. 147-156.

Cox, D. R. (1972) 'Regression Models and Life-Tables', *Journal of the Royal Statistical Society: Series B (Methodological),* 34(2), pp. 187-202.

Cox, T. F. (2022) *Medical Statistics for Cancer Studies.* Chapman and Hall/CRC.

Cruz, J. A. and Wishart, D. S. (2006) 'Applications of Machine Learning in Cancer Prediction and Prognosis', *Cancer Informatics,* 2, pp. 117693510600200030.
Csordas, K., Hegyi, M., Eipel, O. T., Muller, J., Erdelyi, D. J. and Kovacs, G. T. (2013) 'Comparison of pharmacokinetics and toxicity after high-dose methotrexate treatments in children with acute lymphoblastic leukemia', *Anti-Cancer Drugs,* 24(2).
Cuocolo, R., Caruso, M., Perillo, T., Ugga, L. and Petretta, M. (2020) 'Machine Learning in oncology: A clinical appraisal', *Cancer Letters,* 481, pp. 55-62.
Das, P. K., D. V, A., Meher, S., Panda, R. and Abraham, A. (2022) 'A Systematic Review on Recent Advancements in Deep and Machine Learning Based Detection and Classification of Acute Lymphoblastic Leukemia', *IEEE Access,* 10, pp. 81741-81763.
Das, P. K., Pradhan, A. and Meher, S. 'Detection of Acute Lymphoblastic Leukemia Using Machine Learning Techniques'. *Machine Learning, Deep Learning and Computational Intelligence for Wireless Communication*, Singapore, 2021//: Springer Singapore, 425-437.
Dawson, D. V., Blanchette, D. R. and Pihlstrom, B. L. (2021) '13 - Application of Biostatistics in Dental Public Health', in Mascarenhas, A.K., Okunseri, C. and Dye, B.A. (eds.) *Burt and Eklund's Dentistry, Dental Practice, and the Community (Seventh Edition)*. St. Louis: W.B. Saunders, pp. 131-153.
Dekker, S. E., Rea, D., Cayuela, J.-M., Arnhardt, I., Leonard, J. and Heuser, M. (2023) 'Using Measurable Residual Disease to Optimize Management of AML, ALL, and Chronic Myeloid Leukemia', *American Society of Clinical Oncology Educational Book,* (43), pp. e390010.
Dietterich, T. G. 'Ensemble Methods in Machine Learning'. *Multiple Classifier Systems*, Berlin, Heidelberg, 2000//: Springer Berlin Heidelberg, 1-15.
Dixon, S. B., Chen, Y., Yasui, Y., Pui, C.-H., Hunger, S. P., Silverman, L. B., Ness, K. K., Green, D. M., Howell, R. M., Leisenring, W. M., Kadan-Lottick, N. S., Krull, K. R., Oeffinger, K. C., Neglia, J. P., Mertens, A. C., Hudson, M. M., Robison, L. L., Armstrong, G. T. and Nathan, P. C. (2020) 'Reduced Morbidity and Mortality in Survivors of Childhood Acute Lymphoblastic Leukemia: A Report From the Childhood Cancer Survivor Study', *Journal of Clinical Oncology,* 38(29), pp. 3418-3429.
Donadieu, J., Auclerc Mf Fau - Baruchel, A., Baruchel A Fau - Perel, Y., Perel Y Fau - Bordigoni, P., Bordigoni P Fau - Landman-Parker, J., Landman-Parker J Fau - Leblanc, T., Leblanc T Fau - Cornu, G., Cornu G Fau - Sommelet, D., Sommelet D Fau - Leverger, G., Leverger G Fau - Schaison, G., Schaison G Fau - Hill, C. and Hill, C. (2000) 'Prognostic study of continuous variables (white blood cell count, peripheral blast cell count, haemoglobin level, platelet count and age) in childhood acute lymphoblastic leukaemia. Analysis Of a population of 1545 children treated by the French Acute Lymphoblastic Leukaemia Group (FRALLE)', *British Journal of Cancer,* (0007-0920 (Print)).
Duffield, A. S., Mullighan, C. G. and Borowitz, M. J. (2023) 'International Consensus Classification of acute lymphoblastic leukemia/lymphoma', *Virchows Archiv,* 482(1), pp. 11-26.

Eckardt, J.-N., Hahn, W., Röllig, C., Stasik, S., Platzbecker, U., Müller-Tidow, C., Serve, H., Baldus, C. D., Schliemann, C., Schäfer-Eckart, K., Hanoun, M., Kaufmann, M., Burchert, A., Thiede, C., Schetelig, J., Sedlmayr, M., Bornhäuser, M., Wolfien, M. and Middeke, J. M. (2024) 'Mimicking clinical trials with synthetic acute myeloid leukemia patients using generative artificial intelligence', *npj Digital Medicine,* 7(1), pp. 76.

Eden, O. B. (1995) 'Central Nervous System-Directed Therapy in Acute Lymphoblastic Leukemia', *Pediatric Hematology and Oncology,* 12(6), pp. 525-530.

Eden, O. B., Harrison, G., Richards, S., Lilleyman, J. S., Bailey, C. C., Chessells, J. M., Hann, I. M., Hill, F. G. H., Gibson, B. E. S. and on behalf of the Medical Research Council Childhood Leukaemia Working, P. (2000) 'Long-term follow-up of the United Kingdom Medical Research Council protocols for childhood acute lymphoblastic leukaemia, 1980–1997', *Leukemia,* 14(12), pp. 2307-2320.

Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J., Hermsen, M., Manson, Q. F., Balkenhol, M., Geessink, O., Stathonikos, N., van Dijk, M. C., Bult, P., Beca, F., Beck, A. H., Wang, D., Khosla, A., Gargeya, R., Irshad, H., Zhong, A., Dou, Q., Li, Q., Chen, H., Lin, H. J., Heng, P. A., Haß, C., Bruni, E., Wong, Q., Halici, U., Öner, M., Cetin-Atalay, R., Berseth, M., Khvatkov, V., Vylegzhanin, A., Kraus, O., Shaban, M., Rajpoot, N., Awan, R., Sirinukunwattana, K., Qaiser, T., Tsang, Y. W., Tellez, D., Annuscheit, J., Hufnagl, P., Valkonen, M., Kartasalo, K., Latonen, L., Ruusuvuori, P., Liimatainen, K., Albarqouni, S., Mungal, B., George, A., Demirci, S., Navab, N., Watanabe, S., Seno, S., Takenaka, Y., Matsuda, H., Ahmady Phoulady, H., Kovalev, V., Kalinovsky, A., Liauchuk, V., Bueno, G., Fernandez-Carrobles, M. M., Serrano, I., Deniz, O., Racoceanu, D. and Venâncio, R. (2017) 'Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer', *Jama,* 318(22), pp. 2199-2210.

El Naqa, I. and Murphy, M. J. (2015) 'What Is Machine Learning?', in El Naqa, I., Li, R. and Murphy, M.J. (eds.) *Machine Learning in Radiation Oncology: Theory and Applications*. Cham: Springer International Publishing, pp. 3-11.

Elemento, O., Leslie, C., Lundin, J. and Tourassi, G. (2021) 'Artificial intelligence in cancer research, diagnosis and therapy', *Nature Reviews Cancer,* 21(12), pp. 747-752.

Elmore, L. W., Greer, S. F., Daniels, E. C., Saxe, C. C., Melner, M. H., Krawiec, G. M., Cance, W. G. and Phelps, W. C. (2021) 'Blueprint for cancer research: Critical gaps and opportunities', *CA: A Cancer Journal for Clinicians,* 71(2), pp. 107-139.

Enshaei, A., Martinez Elicegui, J., Anguiano, E., Gibson, J., Lawal, S., Ampatzidou, M., Doubek, M., Fielding, A. K., La Sala, E., Middleton, E., Rijneveld, A. W., Turki, A. T., Zimmermann, M., Vora, A. and Moorman, A. V. (2023) 'Real-world evaluation of UK high hyperdiploidy profile using a large cohort of patients provided by HARMONY data platform', *Leukemia,* 37(12), pp. 2493-2496.

Enshaei, A., O'Connor, D., Bartram, J., Hancock, J., Harrison, C. J., Hough, R., Samarasinghe, S., den Boer, M. L., Boer, J. M., de Groot-Kruseman, H. A., Marquart, H. V., Noren-Nystrom, U., Schmiegelow, K., Schwab, C., Horstmann, M. A., Escherich, G., Heyman, M., Pieters, R., Vora, A., Moppett, J. and Moorman, A. V. (2020) 'A validated

novel continuous prognostic index to deliver stratified medicine in pediatric acute lymphoblastic leukemia', *Blood,* 135(17), pp. 1438-1446.
Enshaei, A., Schwab, C. J., Konn, Z. J., Mitchell, C. D., Kinsey, S. E., Wade, R., Vora, A., Harrison, C. J. and Moorman, A. V. (2013) 'Long-term follow-up of ETV6–RUNX1 ALL reveals that NCI risk, rather than secondary genetic abnormalities, is the key risk factor', *Leukemia,* 27(11), pp. 2256-2259.
Enshaei, A., Vora, A., Harrison, C. J., Moppett, J. and Moorman, A. V. (2021) 'Defining low-risk high hyperdiploidy in patients with paediatric acute lymphoblastic leukaemia: a retrospective analysis of data from the UKALL97/99 and UKALL2003 clinical trials', *The Lancet Haematology,* 8(11), pp. e828-e839.
Erdmann, F., Frederiksen, L. E., Bonaventure, A., Mader, L., Hasle, H., Robison, L. L. and Winther, J. F. (2021) 'Childhood cancer: Survival, treatment modalities, late effects and improvements over time', *Cancer Epidemiology,* 71, pp. 101733.
Esparza, S. D. and Sakamoto, K. M. (2005) 'Topics in pediatric leukemia--acute lymphoblastic leukemia', *Medscape General Medicine,* (1531-0132 (Electronic)).
Esposito, C., Landrum, G. A., Schneider, N., Stiefl, N. and Riniker, S. (2021) 'GHOST: Adjusting the Decision Threshold to Handle Imbalanced Data in Machine Learning', *Journal of Chemical Information and Modeling,* 61(6), pp. 2623-2640.
Esposito, F., Malerba, D., Semeraro, G. and Kay, J. (1997) 'A comparative analysis of methods for pruning decision trees', *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 19(5), pp. 476-491.
Essig, S., Li, Q., Chen, Y., Hitzler, J., Leisenring, W., Greenberg, M., Sklar, C., Hudson, M. M., Armstrong, G. T., Krull, K. R., Neglia, J. P., Oeffinger, K. C., Robison, L. L., Kuehni, C. E., Yasui, Y. and Nathan, P. C. (2014) 'Risk of late effects of treatment in children newly diagnosed with standard-risk acute lymphoblastic leukaemia: a report from the Childhood Cancer Survivor Study cohort', *The Lancet Oncology,* 15(8), pp. 841-851.
Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S. and Dean, J. (2019) 'A guide to deep learning in healthcare', *Nature Medicine,* 25(1), pp. 24-29.
Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B. and Herrera, F. (2018a) 'Algorithm-Level Approaches', in Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B. and Herrera, F. (eds.) *Learning from Imbalanced Data Sets*. Cham: Springer International Publishing, pp. 123-146.
Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B. and Herrera, F. (2018b) *Learning from imbalanced data sets.* Springer.
Ferroni, P., Zanzotto, F. M., Riondino, S., Scarpato, N., Guadagni, F. and Roselli, M. (2019) 'Breast Cancer Prognosis Using a Machine Learning Approach', *Cancers*, 11(3). DOI: 10.3390/cancers11030328.
Finkelstein, Y., Zevin, S., Raikhlin-Eisenkraft, B. and Bentur, Y. (2005) 'Intrathecal methotrexate neurotoxicity: clinical correlates and antidotal treatment', *Environ Toxicol Pharmacol,* 19(3), pp. 721-5.

Fišer, K., Sieger, T., Schumich, A., Wood, B., Irving, J., Mejstříková, E. and Dworzak, M. N. (2012) 'Detection and monitoring of normal and leukemic cell populations with hierarchical clustering of flow cytometry data', *Cytometry Part A,* 81A(1), pp. 25-34.
Follin, C. (2019) 'Endocrinopathy After Childhood Cancer Treatment', in Llahana, S., Follin, C., Yedinak, C. and Grossman, A. (eds.) *Advanced Practice in Endocrinology Nursing*. Cham: Springer International Publishing, pp. 1133-1147.
Forestier, E., Heyman, M., Andersen, M. K., Autio, K., Blennow, E., Borgström, G., Golovleva, I., Heim, S., Heinonen, K., Hovland, R., Johannsson, J. H., Kerndrup, G., Nordgren, A., Rosenquist, R., Swolin, B., Johansson, B., the Nordic Society of Paediatric Haematology, O. t. S. C. L. S. G. and the, N. L. C. S. G. (2008) 'Outcome of ETV6/RUNX1-positive childhood acute lymphoblastic leukaemia in the NOPHO-ALL-1992 protocol: frequent late relapses but good overall survival', *British Journal of Haematology,* 140(6), pp. 665-672.
Freedman, L. S. (1982) 'Tables of the number of patients required in clinical trials using the logrank test', *Statistics in Medicine,* 1(2), pp. 121-129.
Freund, Y. (1995) 'Boosting a weak learning algorithm by majority', *Information and computation,* 121(2), pp. 256-285.
Freund, Y. and Schapire, R. E. 'Experiments with a new boosting algorithm'. 1996: Citeseer, 148-156.
Friedman, D. L. and Meadows, A. T. (2002) 'Late effects of childhood cancer therapy', *Pediatric Clinics of North America,* 49(5), pp. 1083-1106.
Friedmann, A. M. and Weinstein, H. J. (2000) 'The Role of Prognostic Features in the Treatment of Childhood Acute Lymphoblastic Leukemia', *The Oncologist,* 5(4), pp. 321-328.
Frost, B.-M., Forestier, E., Gustafsson, G. r., Nygren, P., Hellebostad, M., Jonsson, O. G., Kanerva, J., Schmiegelow, K., Larsson, R., Lönnerholm, G., for the Nordic Society for Paediatric, H. and Oncology (2004) 'Translocation t(12;21) is related to in vitro cellular drug sensitivity to doxorubicin and etoposide in childhood acute lymphoblastic leukemia', *Blood,* 104(8), pp. 2452-2457.
Galathiya, A. S., Ganatra, A. P. and Bhensdadia, C. K. (2012) 'Improved decision tree induction algorithm with feature selection, cross validation, model complexity and reduced error pruning', *International Journal of Computer Science and Information Technologies,* 3(2), pp. 3427-3431.
Gandemer, V., Auclerc, M.-F., Petit, A., Brethon, B., Ballerini, P., Michel, G., Perel, Y., Auvrignon, A., Mechinaud, F., Schmitt, C., Schneider, P., Vermylen, C., Lejars, O., Leblanc, T., Gabert, J., Macintyre, E., Cayuela, J.-M., Leverger, G. and Baruchel, A. (2009) 'Excellent Prognosis of Children with ETV6-RUNX1 Positive (+) Acute Lymphoblastic Leukemia (ALL) in the FRALLE 2000 Protocol', *Blood,* 114(22), pp. 1628.
Geetha, T. V. and Sendhilkumar, S. (2023) *Machine Learning: Concepts, Techniques and Applications.* Chapman and Hall/CRC.
Goldberg, J. M., Silverman, L. B., Levy, D. E., Dalton, V. K., Gelber, R. D., Lehmann, L., Cohen, H. J., Sallan, S. E. and Asselin, B. L. (2003) 'Childhood T-Cell Acute Lymphoblastic Leukemia: The Dana-Farber Cancer Institute Acute Lymphoblastic

Leukemia Consortium Experience', *Journal of Clinical Oncology,* 21(19), pp. 3616-3622.
Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep Learning.* MIT Press.
Goodwin, P. M. (2023) 'Chemotherapy De-Escalation Factors in Acute Lymphoblastic Leukemia', *Oncology Times,* 45(5).
Greaves, M. (2018) 'A causal mechanism for childhood acute lymphoblastic leukaemia', *Nature Reviews Cancer,* 18(8), pp. 471-484.
Groninger, E., Meeuwsen-de Boer, T., Koopmans, P., Uges, D., Sluiter, W., Veerman, A., Kamps, W. and de Graaf, S. (2002) 'Pharmacokinetics of Vincristine Monotherapy in Childhood Acute Lymphoblastic Leukemia', *Pediatric Research,* 52(1), pp. 113-118.
Guo, S. (2010) *Survival analysis.* Oxford University Press.
Guolla, L., Breitbart, S., Foroutan, F., Thabane, L., Loh, M. L., Teachey, D. T., Raetz, E. A. and Gupta, S. (2023) 'Impact of vincristine-steroid pulses during maintenance for B-cell pediatric ALL: a systematic review and meta-analysis', *Blood,* 141(24), pp. 2944-2954.
Gupta, S., Teachey, D. T., Chen, Z., Rabin, K. R., Dunsmore, K. P., Larsen, E. C., Maloney, K. W., Mattano Jr, L. A., Winter, S. S., Carroll, A. J., Heerema, N. A., Borowitz, M. J., Wood, B. L., Carroll, W. L., Raetz, E. A., Winick, N. J., Loh, M. L., Hunger, S. P. and Devidas, M. (2022) 'Sex-based disparities in outcome in pediatric acute lymphoblastic leukemia: a Children's Oncology Group report', *Cancer,* 128(9), pp. 1863-1870.
Gupta, S., Wang, C., Raetz, E. A., Schore, R., Salzer, W. L., Larsen, E. C., Maloney, K. W., Mattano, L. A., Carroll, W. L., Winick, N. J., Hunger, S. P., Loh, M. L. and Devidas, M. (2020) 'Impact of Asparaginase Discontinuation on Outcome in Childhood Acute Lymphoblastic Leukemia: A Report From the Children's Oncology Group', *Journal of Clinical Oncology,* 38(17), pp. 1897-1905.
Haas, O. A. and Borkhardt, A. (2022) 'Hyperdiploidy: the longest known, most prevalent, and most enigmatic form of acute lymphoblastic leukemia in children', *Leukemia,* 36(12), pp. 2769-2783.
Hakeem, A., Shiekh, A. A., Bhat, G. M. and Lone, A. R. (2014) 'Prognostification of ALL by Cytogenetics', *Indian journal of hematology & blood transfusion : an official journal of Indian Society of Hematology and Blood Transfusion,* (0971-4502 (Print)).
Hann, I., Vora, A., Harrison, G., Harrison, C., Eden, O., Hill, F., Gibson, B., Richards, S. and Leukaemia, t. U. M. R. C. s. W. P. o. C. (2001) 'Determinants of outcome after intensified therapy of childhood lymphoblastic leukaemia: results from Medical Research Council United Kingdom acute lymphoblastic leukaemia XI protocol', *British Journal of Haematology,* 113(1), pp. 103-114.
Hann, I., Vora, A., Richards, S., Hill, F., Gibson, B., Lilleyman, J., Kinsey, S., Mitchell, C., Eden, O. B. and on behalf of the, U. K. M. R. C. s. W. P. o. C. L. (2000) 'Benefit of intensified treatment for all children with acute lymphoblastic leukaemia: results from MRC UKALL XI and MRC ALL97 randomised trials', *Leukemia,* 14(3), pp. 356-363.
Harris, M. B., Shuster, J. J., Carroll, A., Look, A. T., Borowitz, M. J., Crist, W. M., Nitschke, R., Pullen, J., Steuber, C. P. and Land, V. J. (1992) 'Trisomy of leukemic cell chromosomes 4 and 10 identifies children with B-progenitor cell acute lymphoblastic leukemia with a

very low risk of treatment failure: a Pediatric Oncology Group study', *Blood,* 79(12), pp. 3316-24.
Harrison, C. J. (2001) 'Acute lymphoblastic leukaemia', *Best Practice & Research Clinical Haematology,* 14(3), pp. 593-607.
Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L. and Bauder, R. A. (2019) 'Severely imbalanced Big Data challenges: investigating data sampling approaches', *Journal of Big Data,* 6(1), pp. 107.
Hasib, K. M., Iqbal, M. S., Shah, F. M., Al Mahmud, J., Popel, M. H., Showrov, M. I. H., Ahmed, S. and Rahman, O. (2020) 'A Survey of Methods for Managing the Classification and Solution of Data Imbalance Problem', *Journal of Computer Science,* 16(11).
Hastings, C., Gaynon Ps Fau - Nachman, J. B., Nachman Jb Fau - Sather, H. N., Sather Hn Fau - Lu, X., Lu X Fau - Devidas, M., Devidas M Fau - Seibel, N. L. and Seibel, N. L. (2014) 'Increased post-induction intensification improves outcome in children and adolescents with a markedly elevated white blood cell count (≥200 × 10(9) /l) with T cell acute lymphoblastic leukaemia but not B cell disease: a report from the Children's Oncology Group', *British Journal of Haematology,* (1365-2141 (Electronic)).
Hayashi, H. A.-O. X., Makimoto, A. A.-O. and Yuza, Y. (2024) 'Treatment of Pediatric Acute Lymphoblastic Leukemia: A Historical Perspective. LID - 10.3390/cancers16040723 [doi] LID - 723', *Cancers (Basal),,* (2072-6694 (Print)).
He, H. and Garcia, E. A. (2009) 'Learning from Imbalanced Data', *IEEE Transactions on Knowledge and Data Engineering,* 21(9), pp. 1263-1284.
Heilmann, J., Vieth, S., Möricke, A., Attarbaschi, A., Barbaric, D., Bodmer, N., Colombini, A., Dalla-Pozza, L., Elitzur, S., Izraeli, S., Mann, G., Niggli, F., Silvestri, D., Stary, J., Rizzari, C., Valsecchi, M. G., Zapotocka, E., Zimmermann, M., Cario, G., Schrappe, M. and Conter, V. (2023) 'Effect of two additional doses of intrathecal methotrexate during induction therapy on serious infectious toxicity in pediatric patients with acute lymphoblastic leukemia', *Haematologica,* 108(12), pp. 3278-3286.
Heim, S. and Mitelman, F. (2009) *Cancer cytogenetics: chromosomal and molecular genetic aberrations of tumor cells.* John Wiley & Sons.
Hinze, L., Möricke, A., Zimmermann, M., Junk, S., Cario, G., Dagdan, E., Kratz, C. P., Conter, V., Schrappe, M. and Stanulla, M. (2017) 'Prognostic impact of IKZF1 deletions in association with vincristine–dexamethasone pulses during maintenance treatment of childhood acute lymphoblastic leukemia on trial ALL-BFM 95', *Leukemia,* 31(8), pp. 1840-1842.
Hoo, Z. H., Candlish, J. and Teare, D. (2017) 'What is an ROC curve?', *Emergency Medicine Journal,* 34(6), pp. 357.
Hosmer Jr, D. W., Lemeshow, S. and Sturdivant, R. X. (2013) *Applied logistic regression.* John Wiley & Sons.
Howard, M. R. and Hamilton, P. J. (2013) *Haematology.* Philadelphia, UNITED KINGDOM: Elsevier Health Sciences.
Howard, S. C. and Pui, C.-H. (2002) 'Endocrine complications in pediatric patients with acute lymphoblastic leukemia', *Blood Reviews,* 16(4), pp. 225-243.

Hryniuk, W. and Bush, H. (1984) 'The importance of dose intensity in chemotherapy of metastatic breast cancer', *Journal of Clinical Oncology,* 2(11), pp. 1281-1288.
Hryniuk, W., Frei, E. and Wright, F. A. (1998) 'A single scale for comparing dose-intensity of all chemotherapy regimens in breast cancer: summation dose-intensity', *Journal of Clinical Oncology,* 16(9), pp. 3137-3147.
Hryniuk, W. and Levine, M. N. (1986) 'Analysis of dose intensity for adjuvant chemotherapy trials in stage II breast cancer', *Journal of Clinical Oncology,* 4(8), pp. 1162-1170.
Huang, W., Song, G., Li, M., Hu, W. and Xie, K. 'Adaptive Weight Optimization for Classification of Imbalanced Data'. *Intelligence Science and Big Data Engineering*, Berlin, Heidelberg, 2013//: Springer Berlin Heidelberg, 546-553.
Huang, X., Li, Y., Zhang, J., Yan, L., Zhao, H., Ding, L., Bhatara, S., Yang, X., Yoshimura, S., Yang, W., Karol, S. E., Inaba, H., Mullighan, C., Litzow, M., Zhu, X., Zhang, Y., Stock, W., Jain, N., Jabbour, E., Kornblau, S. M., Konopleva, M., Pui, C.-H., Paietta, E., Evans, W., Yu, J. and Yang, J. J. (2024) 'Single-cell systems pharmacology identifies development-driven drug response and combination therapy in B cell acute lymphoblastic leukemia', *Cancer Cell,* 42(4), pp. 552-567.e6.
Hunger, S. P., Loh, M. L., Whitlock, J. A., Winick, N. J., Carroll, W. L., Devidas, M., Raetz, E. A. and on behalf of the, C. O. G. A. L. L. C. (2013) 'Children's Oncology Group's 2013 blueprint for research: acute lymphoblastic leukemia', *Pediatric Blood & Cancer,* 60(6), pp. 957-963.
Hunger, S. P. and Mullighan, C. G. (2015) 'Acute Lymphoblastic Leukemia in Children', *New England Journal of Medicine,* 373(16), pp. 1541-1552.
Inaba, H., Greaves, M. and Mullighan, C. G. (2013) 'Acute lymphoblastic leukaemia', *The Lancet,* 381(9881), pp. 1943-1955.
Inaba, H. and Mullighan, C. G. (2020) 'Pediatric acute lymphoblastic leukemia', *Haematologica,* 105(11), pp. 2524-2539.
Inaba, H. and Pui, C.-H. (2010) 'Glucocorticoid use in acute lymphoblastic leukaemia', *The Lancet Oncology,* 11(11), pp. 1096-1106.
Iqbal, T., Elahi, A., Wijns, W. and Shahzad, A. (2022) 'Exploring Unsupervised Machine Learning Classification Methods for Physiological Stress Detection', *Frontiers in Medical Technology,* 4.
Irving, J. A. E. (2016) 'Towards an understanding of the biology and targeted treatment of paediatric relapsed acute lymphoblastic leukaemia', *British Journal of Haematology,* 172(5), pp. 655-666.
Irving, J. A. E., Enshaei, A., Parker, C. A., Sutton, R., Kuiper, R. P., Erhorn, A., Minto, L., Venn, N. C., Law, T., Yu, J., Schwab, C., Davies, R., Matheson, E., Davies, A., Sonneveld, E., den Boer, M. L., Love, S. B., Harrison, C. J., Hoogerbrugge, P. M., Revesz, T., Saha, V. and Moorman, A. V. (2016) 'Integration of genetic and clinical risk factors improves prognostication in relapsed childhood B-cell precursor acute lymphoblastic leukemia', *Blood,* 128(7), pp. 911-922.

Iwamoto, S., Mihara, K., Downing, J. R., Pui, C. H. and Campana, D. (2007) 'Mesenchymal cells regulate the response of acute lymphoblastic leukemia cells to asparaginase', *J Clin Invest,* 117(4), pp. 1049-57.

Jacquy, C., Delepaut, B., Van Daele, S., Vaerman, J. L., Zenebergh, A., Brichard, B., Vermylen, C., Cornu, G. and Martiat, P. (1997) 'A prospective study of minimal residual disease in childhood B-lineage acute lymphoblastic leukaemia: MRD level at the end of induction is a strong predictive factor of relapse', *British Journal of Haematology,* 98(1), pp. 140-146.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2021) *An Introduction to Statistical Learning: with Applications in R.* Springer US.

Jenkins, S. P. (2005) 'Survival analysis', *Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK,* 42, pp. 54-56.

Johnson, J. M. and Khoshgoftaar, T. M. (2019) 'Survey on deep learning with class imbalance', *Journal of Big Data,* 6(1), pp. 27.

Jordan, M. I. and Mitchell, T. M. (2015) 'Machine learning: Trends, perspectives, and prospects', *Science,* 349(6245), pp. 255-260.

Kaplan, E. L. and Meier, P. (1958) 'Nonparametric estimation from incomplete observations', *Journal of the American statistical association,* 53(282), pp. 457-481.

Kaspers, G. J., Smets, L. A., Pieters, R., Van Zantwijk, C. H., Van Wering, E. R. and Veerman, A. J. (1995) 'Favorable prognosis of hyperdiploid common acute lymphoblastic leukemia may be explained by sensitivity to antimetabolites and other drugs: results of an in vitro study', *Blood,* 85(3), pp. 751-6.

Kato, M., Ishimaru, S., Seki, M., Yoshida, K., Shiraishi, Y., Chiba, K., Kakiuchi, N., Sato, Y., Ueno, H., Tanaka, H., Inukai, T., Tomizawa, D., Hasegawa, D., Osumi, T., Arakawa, Y., Aoki, T., Okuya, M., Kaizu, K., Kato, K., Taneyama, Y., Goto, H., Taki, T., Takagi, M., Sanada, M., Koh, K., Takita, J., Miyano, S., Ogawa, S., Ohara, A., Tsuchida, M. and Manabe, A. (2017) 'Long-term outcome of 6-month maintenance chemotherapy for acute lymphoblastic leukemia in children', *Leukemia,* 31(3), pp. 580-584.

Kato, M. and Manabe, A. (2018) 'Treatment and biology of pediatric acute lymphoblastic leukemia', *Pediatrics International,* 60(1), pp. 4-12.

Kawedia, J. D., Kaste, S. C., Pei, D., Panetta, J. C., Cai, X., Cheng, C., Neale, G., Howard, S. C., Evans, W. E., Pui, C.-H. and Relling, M. V. (2011) 'Pharmacokinetic, pharmacodynamic, and pharmacogenetic determinants of osteonecrosis in children with acute lymphoblastic leukemia', *Blood,* 117(8), pp. 2340-2347.

Kebriaei, P., Anastasi, J. and Larson, R. A. (2002) 'Acute lymphoblastic leukaemia: diagnosis and classification', *Best Practice & Research Clinical Haematology,* 15(4), pp. 597-621.

Khosravi, P., Kazemi, E., Imielinski, M., Elemento, O. and Hajirasouliha, I. (2018) 'Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images', *EBioMedicine,* 27, pp. 317-328.

Kim, A. and Jung, I. (2023) 'Optimal selection of resampling methods for imbalanced data with high complexity', *PLoS One,* 18(7), pp. e0288540.

Kim, D. W., Lee, S., Kwon, S., Nam, W., Cha, I.-H. and Kim, H. J. (2019) 'Deep learning-based survival prediction of oral cancer patients', *Scientific Reports,* 9(1), pp. 6994.

Klein, J. P. and Goel, P. K. (2013) 'Survival analysis: state of the art'.
Kloos, R. Q. H., Pieters, R., Jumelet, F. M. V., de Groot-Kruseman, H. A., van den Bos, C. and van der Sluis, I. M. (2020) 'Individualized Asparaginase Dosing in Childhood Acute Lymphoblastic Leukemia', *Journal of Clinical Oncology,* 38(7), pp. 715-724.
Kose, F., Abali, H., Sezer, A., Mertsoylu, H., Disel, U. and Ozyilkan, O. (2009) 'Little dose, huge toxicity: profound hematological toxicity of intrathecal methotrexate', *Leukemia & Lymphoma,* 50(2), pp. 282-283.
Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. and Fotiadis, D. I. (2015) 'Machine learning applications in cancer prognosis and prediction', *Computational and Structural Biotechnology Journal,* 13, pp. 8-17.
Krawczyk, B. (2016) 'Learning from imbalanced data: open challenges and future directions', *Progress in Artificial Intelligence,* 5(4), pp. 221-232.
Kreuger, A., Garwicz, S., Hertz, H., Jonmundsson, C., Latining, M., Lie, S. O., Moe, P. J., Salmi, T. T., Schröder, H., Siimes, M. A., Wesenberg, F., Yssing, M., Åhström, L. and Gustafsson, G. (1991) 'Central Nervous System Disease in Childhood Acute Lymphoblastic Leukemia: Prognostic Factors and Results of Treatment', *Pediatric Hematology and Oncology,* 8(4), pp. 291-299.
Kumar, S., Kaur, P. and Gosain, A. 'A Comprehensive Survey on Ensemble Methods'. *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, 7-9 April 2022, 1-7.
Kunapuli, G. (2023) *Ensemble methods for machine learning.* Simon and Schuster.
Kuruvilla, E. and Kundapura, S. 'Performance Comparison of Machine Learning Algorithms in Groundwater Potability Prediction'. *2022 IEEE 7th International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, 1-3 Dec. 2022, 53-58.
Landier, W., Skinner, R., Wallace, W. H., Hjorth, L., Mulder, R. L., Wong, F. L., Yasui, Y., Bhakta, N., Constine, L. S., Bhatia, S., Kremer, L. C. and Hudson, M. M. (2018) 'Surveillance for Late Effects in Childhood Cancer Survivors', *Journal of Clinical Oncology,* 36(21), pp. 2216-2222.
Langebrake, C., Reinhardt, D. and Ritter, J. (2002) 'Minimising the Long-Term Adverse Effects of Childhood Leukaemia Therapy', *Drug Safety,* 25(15), pp. 1057-1077.
Laningham, F. H., Kun, L. E., Reddick, W. E., Ogg, R. J., Morris, E. B. and Pui, C.-H. (2007) 'Childhood central nervous system leukemia: historical perspectives, current therapy, and acute neurological sequelae', *Neuroradiology,* 49(11), pp. 873-888.
Larkin, T., Kashif, R., Elsayed, A. H., Greer, B., Mangrola, K., Rafiee, R., Nguyen, N., Shastri, V., Horn, B. and Lamba, J. K. (2023) 'Polygenic Pharmacogenomic Markers as Predictors of Toxicity Phenotypes in the Treatment of Acute Lymphoblastic Leukemia: A Single-Center Study', *JCO Precision Oncology,* (7), pp. e2200580.
Lausten-Thomsen, U., Madsen, H. O., Vestergaard, T. R., Hjalgrim, H., Lando, A. and Schmiegelow, K. (2010) 'Increased risk of ALL among premature infants is not explained by increased prevalence of pre-leukemic cell clones', *Blood Cells, Molecules, and Diseases,* 44(3), pp. 188-190.
LeCun, Y., Bengio, Y. and Hinton, G. (2015) 'Deep learning', *Nature,* 521(7553), pp. 436-444.

Lee, J.-c., Kim, J. W., Ahn, S., Kim, H. W., Lee, J., Kim, Y. H., Paik, K.-h., Kim, J. and Hwang, J.-H. (2017) 'Optimal dose reduction of FOLFIRINOX for preserving tumour response in advanced pancreatic cancer: Using cumulative relative dose intensity', *European Journal of Cancer,* 76, pp. 125-133.

Lee, J. W. and Cho, B. (2017) 'Prognostic factors and treatment of pediatric acute lymphoblastic leukemia', *Korean J Pediatr,* 60(5), pp. 129-137.

Lee, S. H. R., Ashcraft, E., Yang, W., Roberts, K. G., Gocho, Y., Rowland, L., Inaba, H., Karol, S. E., Jeha, S., Crews, K. R., Mullighan, C. G., Relling, M. V., Evans, W. E., Cheng, C., Yang, J. J. and Pui, C.-H. (2023a) 'Prognostic and Pharmacotypic Heterogeneity of Hyperdiploidy in Childhood ALL', *Journal of Clinical Oncology,* 41(35), pp. 5422-5432.

Lee, S. H. R., Yang, W., Gocho, Y., John, A., Rowland, L., Smart, B., Williams, H., Maxwell, D., Hunt, J., Yang, W., Crews, K. R., Roberts, K. G., Jeha, S., Cheng, C., Karol, S. E., Relling, M. V., Rosner, G. L., Inaba, H., Mullighan, C. G., Pui, C.-H., Evans, W. E. and Yang, J. J. (2023b) 'Pharmacotypes across the genomic landscape of pediatric acute lymphoblastic leukemia and impact on treatment response', *Nature Medicine,* 29(1), pp. 170-179.

Lejman, M., Kuśmierczuk, K., Bednarz, K., Ostapińska, K. and Zawitkowska, J. (2021) 'Targeted Therapy in the Treatment of Pediatric Acute Lymphoblastic Leukemia—Therapy and Toxicity Mechanisms', *International Journal of Molecular Sciences*, 22(18). DOI: 10.3390/ijms22189827.

Leukaemia and Lymphoma Society (2024) *Measurable Residual Disease (MRD)*.

Levin, L. and Hryniuk, W. M. (1987) 'Dose intensity analysis of chemotherapy regimens in ovarian carcinoma', *Journal of Clinical Oncology,* 5(5), pp. 756-767.

Li, S., Jiang, H. and Pang, W. (2017) 'Joint multiple fully connected convolutional neural network with extreme learning machine for hepatocellular carcinoma nuclei grading', *Comput Biol Med,* 84, pp. 156-167.

Li, X.-Y., Li, J.-Q., Luo, X.-Q., Wu, X.-D., Sun, X., Xu, H.-G., Li, C.-G., Liu, R.-Y., Sun, X.-F., Chen, H.-Q., Lin, Y.-D., Li, C.-k. and Fang, J.-P. (2021) 'Reduced intensity of early intensification does not increase the risk of relapse in children with standard risk acute lymphoblastic leukemia - a multi-centric clinical study of GD-2008-ALL protocol', *BMC Cancer,* 21(1), pp. 59.

Liang, G., Fan, W., Luo, H. and Zhu, X. (2020) 'The emerging roles of artificial intelligence in cancer drug development and precision therapy', *Biomedicine & Pharmacotherapy,* 128, pp. 110255.

Linderman, G. C. and Steinerberger, S. (2019) 'Clustering with t-SNE, Provably', *SIAM Journal on Mathematics of Data Science,* 1(2), pp. 313-332.

Liu, X. (2012) *Survival analysis: models and applications.* John Wiley & Sons.

Liu, Y. Q., Wang, C. and Zhang, L. 'Decision Tree Based Predictive Models for Breast Cancer Survivability on Imbalanced Data'. *2009 3rd International Conference on Bioinformatics and Biomedical Engineering*, 11-13 June 2009, 1-4.

Locatelli, F., Zugmaier, G., Rizzari, C., Morris, J. D., Gruhn, B., Klingebiel, T., Parasole, R., Linderkamp, C., Flotho, C., Petit, A., Micalizzi, C., Mergen, N., Mohammad, A., Kormany, W. N., Eckert, C., Möricke, A., Sartor, M., Hrusak, O., Peters, C., Saha, V.,

Vinti, L. and von Stackelberg, A. (2021) 'Effect of Blinatumomab vs Chemotherapy on Event-Free Survival Among Children With High-risk First-Relapse B-Cell Acute Lymphoblastic Leukemia: A Randomized Clinical Trial', *JAMA,* 325(9), pp. 843-854.

Loh, M. L., Goldwasser, M. A., Silverman, L. B., Poon, W.-M., Vattikuti, S., Cardoso, A., Neuberg, D. S., Shannon, K. M., Sallan, S. E. and Gilliland, D. G. (2006) 'Prospective analysis of TEL/AML1-positive patients treated on Dana-Farber Cancer Institute Consortium Protocol 95-01', *Blood,* 107(11), pp. 4508-4513.

Loke, J. and Kansagra, A. J. (2022) *Fast Facts: Leukemia*: S.Karger AG. Available at: https://doi.org/10.1159/isbn.978-3-318-06949-5.

Lu, W., Fu, D., Kong, X., Huang, Z., Hwang, M., Zhu, Y., Chen, L., Jiang, K., Li, X., Wu, Y., Li, J., Yuan, Y. and Ding, K. (2020) 'FOLFOX treatment response prediction in metastatic or recurrent colorectal cancer patients via machine learning algorithms', *Cancer Med,* 9(4), pp. 1419-1429.

Lustosa de Sousa, D. W., de Almeida Ferreira, F. V., Cavalcante Félix, F. H. and de Oliveira Lopes, M. V. (2015) 'Acute lymphoblastic leukemia in children and adolescents: prognostic factors and analysis of survival', *Revista Brasileira de Hematologia e Hemoterapia,* 37(4), pp. 223-229.

Maclin, R. and Opitz, D. (1997) 'An empirical evaluation of bagging and boosting', *AAAI/IAAI,* 1997, pp. 546-551.

Magnusson, M., Vehtari, A., Jonasson, J. and Andersen, M. 'Leave-One-Out Cross-Validation for Bayesian Model Comparison in Large Data', *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research: PMLR, 341--351.

Mahmood, N., Shahid, S., Bakhshi, T., Riaz, S., Ghufran, H. and Yaqoob, M. (2020) 'Identification of significant risks in pediatric acute lymphoblastic leukemia (ALL) through machine learning (ML) approach', *Medical & Biological Engineering & Computing,* 58(11), pp. 2631-2640.

Malard, F. and Mohty, M. (2020) 'Acute lymphoblastic leukaemia', *The Lancet,* 395(10230), pp. 1146-1162.

Maloney, K. W., Devidas, M., Wang, C., Mattano, L. A., Friedmann, A. M., Buckley, P., Borowitz, M. J., Carroll, A. J., Gastier-Foster, J. M., Heerema, N. A., Kadan-Lottick, N., Loh, M. L., Matloub, Y. H., Marshall, D. T., Stork, L. C., Raetz, E. A., Wood, B., Hunger, S. P., Carroll, W. L. and Winick, N. J. (2019) 'Outcome in Children With Standard-Risk B-Cell Acute Lymphoblastic Leukemia: Results of Children's Oncology Group Trial AALL0331', *Journal of Clinical Oncology,* 38(6), pp. 602-612.

Mody, R., Li, S., Dover, D. C., Sallan, S., Leisenring, W., Oeffinger, K. C., Yasui, Y., Robison, L. L. and Neglia, J. P. (2008) 'Twenty-five–year follow-up among survivors of childhood acute lymphoblastic leukemia: a report from the Childhood Cancer Survivor Study', *Blood,* 111(12), pp. 5515-5523.

Mohamed, W. N. H. W., Salleh, M. N. M. and Omar, A. H. 'A comparative study of Reduced Error Pruning method in decision tree algorithms'. *2012 IEEE International Conference on Control System, Computing and Engineering*, 23-25 Nov. 2012, 392-397.

Mohammed, A. and Kora, R. (2023) 'A comprehensive review on ensemble deep learning: Opportunities and challenges', *Journal of King Saud University - Computer and Information Sciences,* 35(2), pp. 757-774.
Mohammed, M., Khan, M. and Bashier, E. (2016) *Machine Learning: Algorithms and Applications.*
Mohammed, R., Rawashdeh, J. and Abdullah, M. (2020) *Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results.*
Mohapatra, S., Patra, D. and Satpathy, S. (2014) 'An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images', *Neural Computing and Applications,* 24(7), pp. 1887-1904.
Monovich, A. C., Gurumurthy, A. and Ryan, R. J. H. (2024) 'The Diverse Roles of ETV6 Alterations in B-Lymphoblastic Leukemia and Other Hematopoietic Cancers', in Borggrefe, T. and Giaimo, B.D. (eds.) *Transcription factors in blood cell development*. Cham: Springer Nature Switzerland, pp. 291-320.
Moorman, A. V. (2016) 'New and emerging prognostic and predictive genetic biomarkers in B-cell precursor acute lymphoblastic leukemia', *Haematologica,* (1592-8721 (Electronic)).
Moorman, A. V., Antony, G., Wade, R., Butler, E. R., Enshaei, A., Harrison, C. J., Moppett, J., Hough, R., Rowntree, C., Hancock, J., Goulden, N., Samarasinghe, S. and Vora, A. (2022a) 'Time to Cure for Childhood and Young Adult Acute Lymphoblastic Leukemia Is Independent of Early Risk Factors: Long-Term Follow-Up of the UKALL2003 Trial', *Journal of Clinical Oncology,* 40(36), pp. 4228-4239.
Moorman, A. V., Barretta, E., Butler, E. R., Ward, E. J., Twentyman, K., Kirkwood, A. A., Enshaei, A., Schwab, C., Creasey, T., Leongamornlert, D., Papaemmanuil, E., Patrick, P., Clifton-Hadley, L., Patel, B., Menne, T., McMillan, A. K., Harrison, C. J., Rowntree, C. J., Marks, D. I. and Fielding, A. K. (2022b) 'Prognostic impact of chromosomal abnormalities and copy number alterations in adult B-cell precursor acute lymphoblastic leukaemia: a UKALL14 study', *Leukemia,* 36(3), pp. 625-636.
Moorman, A. V., Ensor, H. M., Richards, S. M., Chilton, L., Schwab, C., Kinsey, S. E., Vora, A., Mitchell, C. D. and Harrison, C. J. (2010) 'Prognostic effect of chromosomal abnormalities in childhood B-cell precursor acute lymphoblastic leukaemia: results from the UK Medical Research Council ALL97/99 randomised trial', *The Lancet Oncology,* 11(5), pp. 429-438.
Moorman, A. V., Richards, S. M., Martineau, M., Cheung, K. L., Robinson, H. M., Jalali, G. R., Broadfield, Z. J., Harris, R. L., Taylor, K. E., Gibson, B. E. S., Hann, I. M., Hill, F. G. H., Kinsey, S. E., Eden, T. O. B., Mitchell, C. D., Harrison, C. J. and for the United Kingdom Medical Research Council's Childhood Leukaemia Working, P. (2003) 'Outcome heterogeneity in childhood high-hyperdiploid acute lymphoblastic leukemia', *Blood,* 102(8), pp. 2756-2762.
Moorman, A. V., Robinson, H., Schwab, C., Richards, S. M., Hancock, J., Mitchell, C. D., Goulden, N., Vora, A. and Harrison, C. J. (2013) 'Risk-Directed Treatment Intensification Significantly Reduces the Risk of Relapse Among Children and

Adolescents With Acute Lymphoblastic Leukemia and Intrachromosomal Amplification of Chromosome 21: A Comparison of the MRC ALL97/99 and UKALL2003 Trials', *Journal of Clinical Oncology,* 31(27), pp. 3389-3396.

Moorman, A. V., Schwab, C., Winterman, E., Hancock, J., Castleton, A., Cummins, M., Gibson, B., Goulden, N., Kearns, P., James, B., Kirkwood, A. A., Lancaster, D., Madi, M., McMillan, A., Motwani, J., Norton, A., O'Marcaigh, A., Patrick, K., Bhatnagar, N., Qureshi, A., Richardson, D., Stokley, S., Taylor, G., van Delft, F. W., Moppett, J., Harrison, C. J., Samarasinghe, S. and Vora, A. (2020) 'Adjuvant tyrosine kinase inhibitor therapy improves outcome for children and adolescents with acute lymphoblastic leukaemia who have an ABL-class fusion', *British Journal of Haematology,* 191(5), pp. 844-851.

MoradiAmin, M., Memari, A., Samadzadehaghdam, N., Kermani, S. and Talebi, A. (2016) 'Computer aided detection and classification of acute lymphoblastic leukemia cell subtypes based on microscopic image analysis', *Microsc Res Tech,* 79(10), pp. 908-916.

Möricke, A., Reiter, A., Zimmermann, M., Gadner, H., Stanulla, M., Dördelmann, M., Löning, L., Beier, R., Ludwig, W.-D., Ratei, R., Harbott, J., Boos, J., Mann, G., Niggli, F., Feldges, A., Henze, G., Welte, K., Beck, J.-D., Klingebiel, T., Niemeyer, C., Zintl, F., Bode, U., Urban, C., Wehinger, H., Niethammer, D., Riehm, H., Schrappe, M. and for the German-Austrian-Swiss, A. L. L. B. F. M. S. G. (2008) 'Risk-adjusted therapy of acute lymphoblastic leukemia can decrease treatment burden and improve survival: treatment results of 2169 unselected pediatric and adolescent patients enrolled in the trial ALL-BFM 95', *Blood,* 111(9), pp. 4477-4489.

Mullighan, C. G. (2012) 'The molecular genetic makeup of acute lymphoblastic leukemia', *Hematology,* 2012(1), pp. 389-396.

Muschelli, J. (2020) 'ROC and AUC with a Binary Predictor: a Potentially Misleading Metric', *Journal of Classification,* 37(3), pp. 696-708.

Nahm, F. S. (2022) 'Receiver operating characteristic curve: overview and practical use for clinicians', *kja,* 75(1), pp. 25-36.

Nakatsu, R. T. (2021) 'An Evaluation of Four Resampling Methods Used in Machine Learning Classification', *IEEE Intelligent Systems,* 36(3), pp. 51-57.

Neaga, A. A.-O., Jimbu, L., Mesaros, O., Bota, M., Lazar, D., Cainap, S., Blag, C. A.-O. and Zdrenghea, M. A.-O. X. (2021) 'Why Do Children with Acute Lymphoblastic Leukemia Fare Better Than Adults?', *Cancers (Basal),,* (2072-6694 (Print)).

Ness, K. K. and Gurney, J. G. (2007) 'Adverse Late Effects of Childhood Cancer and Its Treatment on Health and Performance', *Annual Review of Public Health,* 28(1), pp. 279-302.

Nishiwaki, S., Sugiura, I., Koyama, D., Ozawa, Y., Osaki, M., Ishikawa, Y. and Kiyoi, H. (2021) 'Machine learning-aided risk stratification in Philadelphia chromosome-positive acute lymphoblastic leukemia', *Biomarker Research,* 9(1), pp. 13.

Nti, I. K., Nyarko-Boateng, O. and Aning, J. (2021) 'Performance of machine learning algorithms with different K values in K-fold CrossValidation', *International Journal of Information Technology and Computer Science,* 13(6), pp. 61-71.

Nyuytiymbiy, K. (2020) *Parameters and Hyperparameters in Machine Learning and Deep Learning*. Medium. Available at: https://towardsdatascience.com/parameters-and-hyperparameters-aa609601a9ac 2024).

Oeffinger, K. C. and Hudson, M. M. (2004) 'Long-term Complications Following Childhood and Adolescent Cancer: Foundations for Providing Risk-based Health Care for Survivors', *CA: A Cancer Journal for Clinicians,* 54(4), pp. 208-236.

Ongun, G., Halici, U., Leblebicioglu, K., Atalay, V., Beksac, M. and Beksac, S. 'Feature extraction and classification of blood cells for an automated differential blood count system'. *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222)*, 15-19 July 2001, 2461-2466 vol.4.

Onyije, F. M., Olsson, A., Baaken, D., Erdmann, F., Stanulla, M., Wollschläger, D. and Schüz, J. (2022) 'Environmental Risk Factors for Childhood Acute Lymphoblastic Leukemia: An Umbrella Review', *Cancers*, 14(2). DOI: 10.3390/cancers14020382.

Osisanwo, F., Akinsola, J., Awodele, O., Hinmikaiye, J., Olakanmi, O. and Akinjobi, J. (2017) 'Supervised machine learning algorithms: classification and comparison', *International Journal of Computer Trends and Technology (IJCTT),* 48(3), pp. 128-138.

Østergaard, A., Fiocco, M., de Groot-Kruseman, H., Moorman, A. V., Vora, A., Zimmermann, M., Schrappe, M., Biondi, A., Escherich, G., Stary, J., Imai, C., Imamura, T., Heyman, M., Schmiegelow, K. and Pieters, R. (2024) 'ETV6::RUNX1 Acute Lymphoblastic Leukemia: how much therapy is needed for cure?', *Leukemia,* 38(7), pp. 1477-1487.

Pagano, M., Gauvreau, K. and Mattie, H. (2022) *Principles of biostatistics.* Chapman and Hall/CRC.

Pan, L., Liu, G., Lin, F., Zhong, S., Xia, H., Sun, X. and Liang, H. (2017) 'Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia', *Scientific Reports,* 7(1), pp. 7402.

Park, K., Ali, A., Kim, D., An, Y., Kim, M. and Shin, H. (2013) 'Robust predictive model for evaluating breast cancer survivability', *Engineering Applications of Artificial Intelligence,* 26(9), pp. 2194-2205.

Parmar, A., Katariya, R. and Patel, V. 'A Review on Random Forest: An Ensemble Classifier'. *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*, Cham, 2019//: Springer International Publishing, 758-763.

Patel, H., Singh Rajput, D., Thippa Reddy, G., Iwendi, C., Kashif Bashir, A. and Jo, O. (2020) 'A review on classification of imbalanced data for wireless sensor networks', *International Journal of Distributed Sensor Networks,* 16(4), pp. 1550147720916404.

Paulsson, K. (2015) 'High hyperdiploid childhood acute lymphoblastic leukemia: Chromosomal gains as the main driver event', *Molecular & Cellular Oncology,* (2372-3556 (Print)).

Paulsson, K., Forestier, E., Lilljebjörn, H., Heldrup, J., Behrendtz, M., Young, B. D. and Johansson, B. (2010) 'Genetic landscape of high hyperdiploid childhood acute lymphoblastic leukemia', *Proceedings of the National Academy of Sciences,* 107(50), pp. 21719-21724.

Paulsson, K. and Johansson, B. (2009) 'High hyperdiploid childhood acute lymphoblastic leukemia', *Genes, Chromosomes and Cancer,* 48(8), pp. 637-660.
Pedrosa, F., Coustan-Smith, E., Zhou, Y., Cheng, C., Pedrosa, A., Lins, M. M., Pedrosa, M., Lucena-Silva, N., Ramos, A. M. d. L., Vinhas, E., Rivera, G. K., Campana, D. and Ribeiro, R. C. (2020) 'Reduced–dose intensity therapy for pediatric lymphoblastic leukemia: long-term results of the Recife RELLA05 pilot study', *Blood,* 135(17), pp. 1458-1466.
Pérez-Beteta, J., Molina-García, D., Ortiz-Alhambra, J. A., Fernández-Romero, A., Luque, B., Arregui, E., Calvo, M., Borrás, J. M., Meléndez, B., Rodríguez de Lope, Á., Moreno de la Presa, R., Iglesias Bayo, L., Barcia, J. A., Martino, J., Velásquez, C., Asenjo, B., Benavides, M., Herruzo, I., Revert, A., Arana, E. and Pérez-García, V. M. (2018) 'Tumor Surface Regularity at MR Imaging Predicts Survival and Response to Surgery in Patients with Glioblastoma', *Radiology,* 288(1), pp. 218-225.
Pession, A., Valsecchi, M. G., Masera, G., Kamps, W. A., Magyarosy, E., Rizzari, C., van Wering, E. R., Lo Nigro, L., van der Does, A., Locatelli, F., Basso, G. and Aricò, M. (2005) 'Long-Term Results of a Randomized Trial on Extended Use of High Dose l-Asparaginase for Standard Risk Childhood Acute Lymphoblastic Leukemia', *Journal of Clinical Oncology,* 23(28), pp. 7161-7167.
Pieters, R., de Groot-Kruseman, H., Fiocco, M., Verwer, F., Van Overveld, M., Sonneveld, E., van der Velden, V., Beverloo, H. B., Bierings, M., Dors, N., de Haas, V., Hoogerbrugge, P., Van der Sluis, I., Tissing, W., Veening, M., Boer, J. and Den Boer, M. (2023) 'Improved Outcome for ALL by Prolonging Therapy for IKZF1 Deletion and Decreasing Therapy for Other Risk Groups', *Journal of Clinical Oncology,* 41(25), pp. 4130-4142.
Pieters, R., de Groot-Kruseman, H., Van der Velden, V., Fiocco, M., van den Berg, H., de Bont, E., Egeler, R. M., Hoogerbrugge, P., Kaspers, G., Van der Schoot, E., De Haas, V. and Van Dongen, J. (2016) 'Successful Therapy Reduction and Intensification for Childhood Acute Lymphoblastic Leukemia Based on Minimal Residual Disease Monitoring: Study ALL10 From the Dutch Childhood Oncology Group', *Journal of Clinical Oncology,* 34(22), pp. 2591-2601.
Piette, C., Suciu, S., Clappier, E., Bertrand, Y., Drunat, S., Girard, S., Yakouben, K., Plat, G., Dastugue, N., Mazingue, F., Grardel, N., van Roy, N., Uyttebroeck, A., Costa, V., Minckes, O., Sirvent, N., Simon, P., Lutz, P., Ferster, A., Pluchart, C., Poirée, M., Freycon, C., Dresse, M. F., Millot, F., Chantrain, C., van der Werff ten Bosch, J., Norga, K., Gilotay, C., Rohrlich, P. S., Benoit, Y. and Cavé, H. (2018) 'Differential impact of drugs on the outcome of ETV6-RUNX1 positive childhood B-cell precursor acute lymphoblastic leukaemia: results of the EORTC CLG 58881 and 58951 trials', *Leukemia,* 32(1), pp. 244-248.
Pramod, O. (2023) *Decision Trees*. Medium: Medium. Available at: https://medium.com/@ompramod9921/decision-trees-91530198a5a5#:~:text=The%20Gini%20Index%20is%20also,the%20same%20class%20or%20category 2024).

Probst, P., Boulesteix, A.-L. and Bischl, B. (2019) 'Tunability: Importance of hyperparameters of machine learning algorithms', *Journal of Machine Learning Research,* 20(53), pp. 1-32.
Probst, P., Wright, M. N. and Boulesteix, A.-L. (2019) 'Hyperparameters and tuning strategies for random forest', *WIREs Data Mining and Knowledge Discovery,* 9(3), pp. e1301.
Public Health England (2021) *Children, teenagers and young adults UK cancer statistics report*. Available at: http://ncin.org.uk/cancer_type_and_topic_specific_work/cancer_type_specific_work/cancer_in_children_teenagers_and_young_adults/ 2024).
Pui, C.-H. (2006) 'Central Nervous System Disease in Acute Lymphoblastic Leukemia: Prophylaxis and Treatment', *Hematology,* 2006(1), pp. 142-146.
Pui, C.-H. (2012) *Childhood leukemias.* cambridge university press.
Pui, C.-H., Cheng, C., Leung, W., Rai Shesh, N., Rivera Gaston, K., Sandlund John, T., Ribeiro Raul, C., Relling Mary, V., Kun Larry, E., Evans William, E. and Hudson Melissa, M. (2003) 'Extended Follow-up of Long-Term Survivors of Childhood Acute Lymphoblastic Leukemia', *New England Journal of Medicine,* 349(7), pp. 640-649.
Pui, C.-H. and Evans William, E. (2006) 'Treatment of Acute Lymphoblastic Leukemia', *New England Journal of Medicine,* 354(2), pp. 166-178.
Pui, C.-H. and Howard, S. C. (2008) 'Current management and challenges of malignant disease in the CNS in paediatric leukaemia', *The Lancet Oncology,* 9(3), pp. 257-268.
Pui, C.-H. and Jeha, S. (2007) 'New therapeutic strategies for the treatment of acute lymphoblastic leukaemia', *Nature Reviews Drug Discovery,* 6(2), pp. 149-165.
Pui, C.-H., Robison, L. L. and Look, A. T. (2008) 'Acute lymphoblastic leukaemia', *The Lancet,* 371(9617), pp. 1030-1043.
Puri, A. and Kumar Gupta, M. (2022) 'Improved Hybrid Bag-Boost Ensemble With K-Means-SMOTE–ENN Technique for Handling Noisy Class Imbalanced Data', *The Computer Journal,* 65(1), pp. 124-138.
Putzu, L., Caocci, G. and Di Ruberto, C. (2014) 'Leucocyte classification for leukaemia detection using image processing techniques', *Artif Intell Med,* 62(3), pp. 179-91.
Rad, A. and Häggström, M. 2009. Haematopoesis.
Rawat, J., Singh, A., Bhadauria, H. S., Virmani, J. and Devgun, J. S. (2017) 'Classification of acute lymphoblastic leukaemia using hybrid hierarchical classifiers', *Multimedia Tools and Applications,* 76(18), pp. 19057-19085.
Rawat, S. S. and Mishra, A. K. 'Review of Methods for Handling Class Imbalance in Classification Problems'. *Data Engineering and Applications*, Singapore, 2024//: Springer Nature Singapore, 3-14.
Redaelli, A., Laskin, B. L., Stephens, J. M., Botteman, M. F. and Pashos, C. L. (2005) 'A systematic literature review of the clinical and epidemiological burden of acute lymphoblastic leukaemia (ALL)', *European Journal of Cancer Care,* 14(1), pp. 53-62.
Rehman, A., Abbas, N., Saba, T., Rahman, S. I. u., Mehmood, Z. and Kolivand, H. (2018) 'Classification of acute lymphoblastic leukemia using deep learning', *Microscopy Research and Technique,* 81(11), pp. 1310-1317.

Reismüller, B., Steiner, M., Pichler, H., Dworzak, M., Urban, C., Meister, B., Schmitt, K., Pötschger, U., König, M., Mann, G., Haas, O. A., Attarbaschi, A. and Austrian, A. L. L. B. F. M. S. G. (2017) 'High hyperdiploid acute lymphoblastic leukemia (ALL)—A 25-year population-based survey of the Austrian ALL-BFM (Berlin-Frankfurt-Münster) Study Group', *Pediatric Blood & Cancer,* 64(6), pp. e26327.

Reta, C., Altamirano, L., Gonzalez, J. A., Diaz-Hernandez, R., Peregrina, H., Olmos, I., Alonso, J. E. and Lobato, R. (2015) 'Segmentation and Classification of Bone Marrow Cells Images Using Contextual Information for Medical Diagnosis of Acute Leukemias', *PLoS One,* 10(6), pp. e0130805.

Richards, S., Pui, C.-H., Gayon, P. and on behalf of the Childhood Acute Lymphoblastic Leukemia Collaborative, G. (2013) 'Systematic review and meta-analysis of randomized trials of central nervous system directed therapy for childhood acute lymphoblastic leukemia', *Pediatric Blood & Cancer,* 60(2), pp. 185-195.

Riehm, H. (1991) 'Results and Significance of Six Randomized Trials in Four Consecutive ALL-BFM Studies', in Kobayashi, N., Akera, T. and Mizutani, S. (eds.) *Childhood Leukemia: Present Problems and Future Prospects: Proceedings of the Second International Symposium on Children's Cancer Tokyo, Japan, December 7–9, 1989*. Boston, MA: Springer US, pp. 135-147.

Rigatti, S. J. (2017) 'Random Forest', *Journal of Insurance Medicine,* 47(1), pp. 31-39.

Roberts, K. G. (2018) 'Genetics and prognosis of ALL in children vs adults', *Hematology Am Soc Hematol Educ Program,* 2018(1), pp. 137-145.

Roel, P., Marc, B. B., Cindy, S. v. d. L., Mathijs, A. S., Onno, R., João, R. M. M., Judith, M. B., Jan, J. C., Rob, P., Monique, L. d. B. and Miranda, B. (2019) 'Autophagy inhibition as a potential future targeted therapy for ETV6-RUNX1-driven B-cell precursor acute lymphoblastic leukemia', *Haematologica,* 104(4), pp. 738-748.

Rogers, J. H., Gupta, R., Reyes, J. M., Gundry, M. C., Medrano, G., Guzman, A., Aguilar, R., Conneely, S. E., Song, T., Johnson, C., Barnes, S., Cristobal, C. D. D., Kurtz, K., Brunetti, L., Goodell, M. A. and Rau, R. E. (2021) 'Modeling IKZF1 lesions in B-ALL reveals distinct chemosensitivity patterns and potential therapeutic vulnerabilities', *Blood Advances,* 5(19), pp. 3876-3890.

Rubnitz, J. E., Wichlan, D., Devidas, M., Shuster, J., Linda, S. B., Kurtzberg, J., Bell, B., Hunger, S. P., Chauvenet, A., Pui, C.-H., Camitta, B. and Pullen, J. (2008) 'Prospective Analysis of TEL Gene Rearrangements in Childhood Acute Lymphoblastic Leukemia: A Children's Oncology Group Study', *Journal of Clinical Oncology,* 26(13), pp. 2186-2191.

Rüchel, N., Jepsen, V. H., Hein, D., Fischer, U., Borkhardt, A. and Gössling, K. L. (2022) 'In Utero Development and Immunosurveillance of B Cell Acute Lymphoblastic Leukemia', *Current Treatment Options in Oncology,* 23(4), pp. 543-561.

Sagi, O. and Rokach, L. (2018) 'Ensemble learning: A survey', *WIREs Data Mining and Knowledge Discovery,* 8(4), pp. e1249.

Sahoo, S. S., Kobow, K., Zhang, J., Buchhalter, J., Dayyani, M., Upadhyaya, D. P., Prantzalos, K., Bhattacharjee, M., Blumcke, I., Wiebe, S. and Lhatoo, S. D. (2022) 'Ontology-based feature engineering in machine learning workflows for heterogeneous epilepsy patient records', *Scientific Reports,* 12(1), pp. 19430.

Saito, T. R., Marc (2024). Available at: https://classeval.wordpress.com/introduction/introduction-to-the-roc-receiver-operating-characteristics-plot/ 2024).
Sallan, S. E., Hitchcock-Bryan, S., Gelber, R., Cassady, J. R., Frei, E., III and Nathan, D. G. (1983) 'Influence of Intensive Asparaginase in the Treatment of Childhood Non-T-Cell Acute Lymphoblastic Leukemia', *Cancer Research,* 43(11), pp. 5601-5607.
Samarasinghe, S., Vora, A., Goulden, N. J., Antony, G. and Moorman, A. V. (2021) 'Ten Year Outcomes of UKALL 2003: A Randomised Clinical Trial of Adjusting Treatment Intensity Based on Minimal Residual Disease', *Blood,* 138, pp. 364.
Samra, B., Jabbour, E., Ravandi, F., Kantarjian, H. and Short, N. J. (2020) 'Evolving therapy of adult acute lymphoblastic leukemia: state-of-the-art treatment and future directions', *Journal of Hematology & Oncology,* 13(1), pp. 70.
Sarker, I. H. (2021) 'Machine Learning: Algorithms, Real-World Applications and Research Directions', *SN Computer Science,* 2(3), pp. 160.
Sasada, T., Liu, Z., Baba, T., Hatano, K. and Kimura, Y. (2020) 'A Resampling Method for Imbalanced Datasets Considering Noise and Overlap', *Procedia Computer Science,* 176, pp. 420-429.
Schapire, R. E. (1990) 'The strength of weak learnability', *Machine learning,* 5, pp. 197-227.
Schmiegelow, K., Attarbaschi, A., Barzilai, S., Escherich, G., Frandsen, T. L., Halsey, C., Hough, R., Jeha, S., Kato, M., Liang, D.-C., Mikkelsen, T. S., Möricke, A., Niinimäki, R., Piette, C., Putti, M. C., Raetz, E., Silverman, L. B., Skinner, R., Tuckuviene, R., van der Sluis, I. and Zapotocka, E. (2016) 'Consensus definitions of 14 severe acute toxic effects for childhood lymphoblastic leukaemia treatment: a Delphi consensus', *The Lancet Oncology,* 17(6), pp. e231-e239.
Schore, R. J., Angiolillo, A. L., Kairalla, J. A., Devidas, M., Rabin, K. R., Zweidler-McKay, P., Borowitz, M. J., Wood, B., Carroll, A. J., Heerema, N. A., Relling, M. V., Hitzler, J., Kadan-Lottick, N. S., Maloney, K., Wang, C., Carroll, W. L., Winick, N. J., Raetz, E. A., Loh, M. L. and Hunger, S. P. (2023) 'Outstanding outcomes with two low intensity regimens in children with low-risk B-ALL: a report from COG AALL0932', *Leukemia,* 37(6), pp. 1375-1378.
Schrappe, M., Bleckmann, K., Zimmermann, M., Biondi, A., Möricke, A., Locatelli, F., Cario, G., Rizzari, C., Attarbaschi, A., Valsecchi, M. G., Bartram, C. R., Barisone, E., Niggli, F., Niemeyer, C., Testi, A. M., Mann, G., Ziino, O., Schäfer, B., Panzer-Grümayer, R., Beier, R., Parasole, R., Göhring, G., Ludwig, W.-D., Casale, F., Schlegel, P.-G., Basso, G. and Conter, V. (2017) 'Reduced-Intensity Delayed Intensification in Standard-Risk Pediatric Acute Lymphoblastic Leukemia Defined by Undetectable Minimal Residual Disease: Results of an International Randomized Trial (AIEOP-BFM ALL 2000)', *Journal of Clinical Oncology,* 36(3), pp. 244-253.
Schrappe, M., Hunger Stephen, P., Pui, C.-H., Saha, V., Gaynon Paul, S., Baruchel, A., Conter, V., Otten, J., Ohara, A., Versluys Anne, B., Escherich, G., Heyman, M., Silverman Lewis, B., Horibe, K., Mann, G., Camitta Bruce, M., Harbott, J., Riehm, H., Richards, S., Devidas, M. and Zimmermann, M. (2012) 'Outcomes after Induction

Failure in Childhood Acute Lymphoblastic Leukemia', *New England Journal of Medicine,* 366(15), pp. 1371-1381.
Schrappe, M., Reiter, A., Ludwig, W.-D., Harbott, J., Zimmermann, M., Hiddemann, W., Niemeyer, C., Henze, G., Feldges, A., Zintl, F., Kornhuber, B., Ritter, J., Welte, K., Gadner, H. and Riehm, H. (2000) 'Improved outcome in childhood acute lymphoblastic leukemia despite reduced use of anthracyclines and cranial radiotherapy: results of trial ALL-BFM 90', *Blood,* 95(11), pp. 3310-3322.
Schwab, C. J., Murdy, D., Butler, E., Enshaei, A., Winterman, E., Cranston, R. E., Ryan, S., Barretta, E., Hawking, Z., Murray, J., Antony, G., Vora, A., Moorman, A. V. and Harrison, C. J. (2022) 'Genetic characterisation of childhood B-other-acute lymphoblastic leukaemia in UK patients by fluorescence in situ hybridisation and Multiplex Ligation-dependent Probe Amplification', *British Journal of Haematology,* 196(3), pp. 753-763.
Sebastian, A. M. and Peter, D. (2022) 'Artificial Intelligence in Cancer Research: Trends, Challenges and Future Directions', *Life*, 12(12). DOI: 10.3390/life12121991.
Seibel, N. L. (2008) 'Treatment of Acute Lymphoblastic Leukemia in Children and Adolescents: Peaks and Pitfalls', *Hematology,* 2008(1), pp. 374-380.
Sen, P. C., Hajra, M. and Ghosh, M. 'Supervised Classification Algorithms in Machine Learning: A Survey and Review'. *Emerging Technology in Modelling and Graphics*, Singapore, 2020//: Springer Singapore, 99-111.
Shafique, S. and Tehsin, S. (2018) 'Acute Lymphoblastic Leukemia Detection and Classification of Its Subtypes Using Pretrained Deep Convolutional Neural Networks', *Technol Cancer Res Treat,* 17, pp. 1533033818802789.
Shamrat, F. M. J. M., Chakraborty, S., Billah, M. M., Das, P., Muna, J. N. and Ranjan, R. 'A comprehensive study on pre-pruning and post-pruning methods of decision tree classification algorithm'. *2021 5th International Conference on Trends in Electronics and Informatics (ICOEI)*, 3-5 June 2021, 1339-1345.
Sharathkumar, A., DeCamillo, D., Bhambhani, K., Cushing, B., Thomas, R., Mohamed, A. N., Ravindranath, Y. and Taub, J. W. (2008) 'Children with hyperdiploid but not triple trisomy (+4,+10,+17) acute lymphoblastic leukemia have an increased incidence of extramedullary relapse on current therapies: A single institution experience', *American Journal of Hematology,* 83(1), pp. 34-40.
Shouval, R., Fein, J. A., Savani, B., Mohty, M. and Nagler, A. (2021) 'Machine learning and artificial intelligence in haematology', *British Journal of Haematology,* 192(2), pp. 239-250.
Shuster, J. J., Falletta, J. M., Pullen, D. J., Crist, W. M., Humphrey, G. B., Dowell, B. L., Wharam, M. D. and Borowitz, M. (1990) 'Prognostic factors in childhood T-cell acute lymphoblastic leukemia: a Pediatric Oncology Group study', *Blood,* 75(1), pp. 166-173.
Sidhom, I., Shaaban, K., Youssef, S. H., Ali, N., Gohar, S., Rashed, W. M., Mehanna, M., Salem, S., Soliman, S., Yassin, D., Mansour, E., Coustan-Smith, E., Ribeiro, R. C. and Rivera, G. K. (2021) 'Reduced-intensity therapy for pediatric lymphoblastic leukemia: impact of residual disease early in remission induction', *Blood,* 137(1), pp. 20-28.

Silverman, L. B., Gelber, R. D., Dalton, V. K., Asselin, B. L., Barr, R. D., Clavell, L. A., Hurwitz, C. A., Moghrabi, A., Samson, Y., Schorin, M. A., Arkin, S., Declerck, L., Cohen, H. J. and Sallan, S. E. (2001) 'Improved outcome for children with acute lymphoblastic leukemia: results of Dana-Farber Consortium Protocol 91-01', *Blood,* 97(5), pp. 1211-1218.

Singh, A., Thakur, N. and Sharma, A. 'A review of supervised machine learning algorithms'. *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 16-18 March 2016, 1310-1315.

Søegaard, S. H., Rostgaard, K., Kamper-Jørgensen, M., Schmiegelow, K. and Hjalgrim, H. (2018) 'Maternal diabetes and risk of childhood acute lymphoblastic leukaemia in the offspring', *British Journal of Cancer,* 118(1), pp. 117-120.

Stanulla, M., Dagdan, E., Zaliova, M., Möricke, A., Palmi, C., Cazzaniga, G., Eckert, C., te Kronnie, G., Bourquin, J.-P., Bornhauser, B., Koehler, R., Bartram, C. R., Ludwig, W.-D., Bleckmann, K., Groeneveld-Krentz, S., Schewe, D., Junk, S. V., Hinze, L., Klein, N., Kratz, C. P., Biondi, A., Borkhardt, A., Kulozik, A., Muckenthaler, M. U., Basso, G., Valsecchi, M. G., Izraeli, S., Petersen, B.-S., Franke, A., Dörge, P., Steinemann, D., Haas, O. A., Panzer-Grümayer, R., Cavé, H., Houlston, R. S., Cario, G., Schrappe, M. and Zimmermann, M. (2018) 'IKZF1plus Defines a New Minimal Residual Disease–Dependent Very-Poor Prognostic Profile in Pediatric B-Cell Precursor Acute Lymphoblastic Leukemia', *Journal of Clinical Oncology,* 36(12), pp. 1240-1249.

Steliarova-Foucher, E., Stiller, C., Lacour, B. and Kaatsch, P. (2005) 'International Classification of Childhood Cancer, third edition', *Cancer,* 103(7), pp. 1457-1467.

Sun, C., Chang, L. and Zhu, X. (2017) 'Pathogenesis of ETV6/RUNX1-positive childhood acute lymphoblastic leukemia and mechanisms underlying its relapse', *Oncotarget,* 8(21), pp. 35445-35459.

Sutcliffe, M. J., Shuster, J. J., Sather, H. N., Camitta, B. M., Pullen, J., Schultz, K. R., Borowitz, M. J., Gaynon, P. S., Carroll, A. J. and Heerema, N. A. (2005) 'High concordance from independent studies by the Children's Cancer Group (CCG) and Pediatric Oncology Group (POG) associating favorable prognosis with combined trisomies 4, 10, and 17 in children with NCI Standard-Risk B-precursor Acute Lymphoblastic Leukemia: a Children's Oncology Group (COG) initiative', *Leukemia,* 19(5), pp. 734-40.

Suthaharan, S. (2016) 'Decision Tree Learning', in Suthaharan, S. (ed.) *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*. Boston, MA: Springer US, pp. 237-269.

Sutton, R., Venn, N. C., Tolisano, J., Bahar, A. Y., Giles, J. E., Ashton, L. J., Teague, L., Rigutto, G., Waters, K., Marshall, G. M., Haber, M., Norris, M. D., Australian and New Zealand Children's Oncology, G. (2009) 'Clinical significance of minimal residual disease at day 15 and at the end of therapy in childhood acute lymphoblastic leukaemia', *British Journal of Haematology,* 146(3), pp. 292-299.

Suyal, M. and Goyal, P. (2022) 'A review on analysis of K-nearest neighbor classification machine learning algorithms based on supervised learning', *International Journal of Engineering Trends and Technology,* 70(7), pp. 43-48.

Swaminathan, S., Klemm, L., Park, E., Papaemmanuil, E., Ford, A., Kweon, S.-M., Trageser, D., Hasselfeld, B., Henke, N. and Mooster, J. (2015) 'Mechanisms of clonal evolution in childhood acute lymphoblastic leukemia', *Nature immunology,* 16(7), pp. 766-774.
Synold, T. W., Relling, M. V., Boyett, J. M., Rivera, G. K., Sandlund, J. T., Mahmoud, H., Crist, W. M., Pui, C. H. and Evans, W. E. (1994) 'Blast cell methotrexate-polyglutamate accumulation in vivo differs by lineage, ploidy, and methotrexate dose in acute lymphoblastic leukemia', *The Journal of Clinical Investigation,* 94(5), pp. 1996-2001.
Tan, H. (2021) 'Machine Learning Algorithm for Classification', *Journal of Physics: Conference Series,* 1994(1), pp. 012016.
Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N. and Asadpour, M. (2020) 'Boosting methods for multi-class imbalanced data classification: an experimental review', *Journal of Big Data,* 7(1), pp. 70.
Teachey, D. T., Hunger, S. P. and Loh, M. L. (2021) 'Optimizing therapy in the modern age: differences in length of maintenance therapy in acute lymphoblastic leukemia', *Blood,* 137(2), pp. 168-177.
Teachey, D. T. and Pui, C.-H. (2019) 'Comparative features and outcomes between paediatric T-cell and B-cell acute lymphoblastic leukaemia', *The Lancet Oncology,* 20(3), pp. e142-e154.
Tebbi, C. K. (2021) 'Etiology of Acute Leukemia: A Review', *Cancers*, 13(9). DOI: 10.3390/cancers13092256.
Thakar, C. T., Shagufta (2022) *Gini Index: Decision Tree, Formula, Calculator, Gini Coefficient in Machine Learning*. QuantInsti: QuantInsti. Available at: https://blog.quantinsti.com/gini-index/ 2024).
Thastrup, M., Duguid, A., Mirian, C., Schmiegelow, K. and Halsey, C. (2022) 'Central nervous system involvement in childhood acute lymphoblastic leukemia: challenges and solutions', *Leukemia,* 36(12), pp. 2751-2768.
Tzanis, G., Katakis, I., Partalas, I. and Vlahavas, I. (2006) 'Modern Applications of Machine Learning'.
van Delft, F. W., Horsley, S., Colman, S., Anderson, K., Bateman, C., Kempski, H., Zuna, J., Eckert, C., Saha, V., Kearney, L., Ford, A. and Greaves, M. (2011) 'Clonal origins of relapse in ETV6-RUNX1 acute lymphoblastic leukemia', *Blood,* 117(23), pp. 6247-6254.
Van der Maaten, L. and Hinton, G. (2008) 'Visualizing data using t-SNE', *Journal of machine learning research,* 9(11).
Van Vlierberghe, P., Pieters, R., Beverloo, H. B. and Meijerink, J. P. P. (2008) 'Molecular-genetic insights in paediatric T-cell acute lymphoblastic leukaemia', *British Journal of Haematology,* 143(2), pp. 153-168.
Varoquaux, G. and Colliot, O. (2023) 'Evaluating Machine Learning Models and Their Diagnostic Value', in Colliot, O. (ed.) *Machine Learning for Brain Disorders*. New York, NY: Springer US, pp. 601-630.
Vary, A., Lebellec, L., Di Fiore, F., Penel, N., Cheymol, C., Rad, E., El Hajbi, F., Lièvre, A., Edeline, J., Bimbai, A. M., Le Deley, M.-C. and Turpin, A. (2021) 'FOLFIRINOX relative dose intensity and disease control in advanced pancreatic adenocarcinoma', *Therapeutic Advances in Medical Oncology,* 13, pp. 17588359211029825.

Vilmer, E., Suciu, S., Ferster, A., Bertrand, Y., Cavé, H., Thyss, A., Benoit, Y., Dastugue, N., Fournier, M., Souillet, G., Manel, A. M., Robert, A., Nelken, B., Millot, F., Lutz, P., Rialland, X., Mechinaud, F., Boutard, P., Behar, C., Chantraine, J. M., Plouvier, E., Laureys, G., Brock, P., Uyttebroeck, A., Margueritte, G., Plantaz, D., Norton, L., Francotte, N., Gyselinck, J., Waterkeyn, C., Solbu, G., Philippe, N. and Otten, J. (2000) 'Long-term results of three randomized trials (58831, 58832, 58881) in childhood acute lymphoblastic leukemia: a CLCG-EORTC report', *Leukemia,* 14(12), pp. 2257-2266.

Viswanathan, A., Kumar, A., Kaushik, P., Thumallapalli, A., Ramachandra, C., Kumari, B., Appaji, L. and Kumar, N. (2021) 'Administration and Toxicity Profile of the Capizzi Interim Maintenance—Retrospective Study from a Tertiary Care Cancer Centre', *Indian Journal of Medical and Paediatric Oncology,* 42.

Vora, A., Andreano, A., Pui, C. H., Hunger, S. P., Schrappe, M., Moericke, A., Biondi, A., Escherich, G., Silverman, L. B., Goulden, N., Taskinen, M., Pieters, R., Horibe, K., Devidas, M., Locatelli, F. and Valsecchi, M. G. (2016) 'Influence of Cranial Radiotherapy on Outcome in Children With Acute Lymphoblastic Leukemia Treated With Contemporary Therapy', *Journal of Clinical Oncology,* (1527-7755 (Electronic)).

Vora, A., Goulden, N., Mitchell, C., Hancock, J., Hough, R., Rowntree, C., Moorman, A. V. and Wade, R. (2014) 'Augmented post-remission therapy for a minimal residual disease-defined high-risk subgroup of children and young people with clinical standard-risk and intermediate-risk acute lymphoblastic leukaemia (UKALL 2003): a randomised controlled trial', *The Lancet Oncology,* 15(8), pp. 809-818.

Vujović, Ž. (2021) 'Classification model evaluation metrics', *International Journal of Advanced Computer Science and Applications,* 12(6), pp. 599-606.

Vuttipittayamongkol, P., Elyan, E. and Petrovski, A. (2021) 'On the class overlap problem in imbalanced data classification', *Knowledge-Based Systems,* 212, pp. 106631.

Wang, X., Yang, W., Weinreb, J., Han, J., Li, Q., Kong, X., Yan, Y., Ke, Z., Luo, B., Liu, T. and Wang, L. (2017) 'Searching for prostate cancer by fully automated magnetic resonance imaging classification: deep learning versus non-deep learning', *Sci Rep,* 7(1), pp. 15415.

Warris, L. T., van den Heuvel-Eibrink, M. M., Aarsen, F. K., Pluijm, S. M. F., Bierings, M. B., van den Bos, C., Zwaan, C. M., Thygesen, H. H., Tissing, W. J. E., Veening, M. A., Pieters, R. and van den Akker, E. L. T. (2016) 'Hydrocortisone as an Intervention for Dexamethasone-Induced Adverse Effects in Pediatric Patients With Acute Lymphoblastic Leukemia: Results of a Double-Blind, Randomized Controlled Trial', *Journal of Clinical Oncology,* 34(19), pp. 2287-2293.

Welvaars, K., Oosterhoff, J. H. F., van den Bekerom, M. P. J., Doornberg, J. N., van Haarst, E. P., Consortium, O. U. and the Machine Learning, C. (2023) 'Implications of resampling data to address the class imbalance problem (IRCIP): an evaluation of impact on performance between classification algorithms in medical data', *JAMIA Open,* 6(2), pp. ooad033.

Whitehead, V. M., Vuchich, M. J., Cooley, L. D., Lauer, S. J., Mahoney, D. H., Shuster, J. J., Payment, C., Koch, P. A., Akabutu, J. J., Bowen, T., Kamen, B. A., Ravindranath, Y.,

Emami, A., Look, A. T., Beardsley, G. P., Pullen, D. J. and Camitta, B. (1998) 'Accumulation of Methotrexate Polyglutamates, Ploidy and Trisomies of Both Chromosomes 4 and 10 in Lymphoblasts from Children with B-Progenitor Cell Acute Lymphoblastic Leukemia: a Pediatric Oncology Group Study', *Leukemia & Lymphoma,* 31(5-6), pp. 507-519.

Williams, L. A., Yang, J. J., Hirsch, B. A., Marcotte, E. L. and Spector, L. G. (2019) 'Is There Etiologic Heterogeneity between Subtypes of Childhood Acute Lymphoblastic Leukemia? A Review of Variation in Risk by Subtype', *Cancer Epidemiology, Biomarkers & Prevention,* 28(5), pp. 846-856.

Woerden, N. L. R.-v., Pieters, R., Loonen, A. H., Hubeek, I., van Drunen, E., Beverloo, H. B., Slater, R. M., Harbott, J., Seyfarth, J., van Wering, E. R., Hählen, K., Schmiegelow, K., Janka-Schaub, G. E. and Veerman, A. J. P. (2000) 'TEL/AML1 gene fusion is related to in vitro drug sensitivity forl-asparaginase in childhood acute lymphoblastic leukemia', *Blood,* 96(3), pp. 1094-1099.

Wong, T.-T. (2015) 'Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation', *Pattern Recognition,* 48(9), pp. 2839-2846.

Woodward, A. A., Urbanowicz, R. J., Naj, A. C. and Moore, J. H. (2022) 'Genetic heterogeneity: Challenges, impacts, and methods through an associative lens', *Genet Epidemiol,* 46(8), pp. 555-571.

Woodward, E. L., Yang, M., Moura-Castro, L. H., van den Bos, H., Gunnarsson, R., Olsson-Arvidsson, L., Spierings, D. C. J., Castor, A., Duployez, N., Zaliova, M., Zuna, J., Johansson, B., Foijer, F. and Paulsson, K. (2023) 'Clonal origin and development of high hyperdiploidy in childhood acute lymphoblastic leukaemia', *Nature Communications,* 14(1), pp. 1658.

Wu, C. and Li, W. (2018) 'Genomics and pharmacogenomics of pediatric acute lymphoblastic leukemia', *Critical Reviews in Oncology/Hematology,* 126, pp. 100-111.

Xiao, H. and Wang, Y. (2019) 'A systematical approach to classification problems with feature space heterogeneity', *Kybernetes,* 48(9), pp. 2006-2029.

Yala, A., Lehman, C., Schuster, T., Portnoi, T. and Barzilay, R. (2019) 'A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction', *Radiology,* 292(1), pp. 60-66.

Yang, J. J., Cheng, C., Yang, W., Pei, D., Cao, X., Fan, Y., Pounds, S. B., Neale, G., Treviño, L. R., French, D., Campana, D., Downing, J. R., Evans, W. E., Pui, C.-H., Devidas, M., Bowman, W. P., Camitta, B. M., Willman, C. L., Davies, S. M., Borowitz, M. J., Carroll, W. L., Hunger, S. P. and Relling, M. V. (2009) 'Genome-wide Interrogation of Germline Genetic Variation Associated With Treatment Response in Childhood Acute Lymphoblastic Leukemia', *JAMA,* 301(4), pp. 393-403.

Yang, L., Panetta, J. C., Cai, X., Yang, W., Pei, D., Cheng, C., Kornegay, N., Pui, C.-H. and Relling, M. V. (2008) 'Asparaginase May Influence Dexamethasone Pharmacokinetics in Acute Lymphoblastic Leukemia', *Journal of Clinical Oncology,* 26(12), pp. 1932-1939.

Yoshimura, S., Li, Z., Gocho, Y., Yang, W., Crews, K. R., Lee, S. H. R., Roberts, K. G., Mullighan, C. G., Relling, M. V., Yu, J., Yeoh, A. E. J., Loh, M. L., Saygin, C., Litzow, M. R.,

Jeha, S., Karol, S. E., Inaba, H., Pui, C.-H., Konopleva, M., Jain, N., Stock, W., Paietta, E., Jabbour, E., Kornblau, S. M., Evans, W. E. and Yang, J. J. (2024) 'Impact of Age on Pharmacogenomics and Treatment Outcomes of B-Cell Acute Lymphoblastic Leukemia', *Journal of Clinical Oncology,* 0(0), pp. JCO.24.00500.

Zhan, M., Chen, Z.-b., Ding, C.-c., Qu, Q., Wang, G.-q., Liu, S. and Wen, F.-q. (2021) 'Machine learning to predict high-dose methotrexate-related neutropenia and fever in children with B-cell acute lymphoblastic leukemia', *Leukemia & Lymphoma,* 62(10), pp. 2502-2513.

Zhang, B., Shi, H. and Wang, H. (2023) 'Machine Learning and AI in Cancer Prognosis, Prediction, and Treatment Selection: A Critical Approach', *J Multidiscip Healthc,* 16, pp. 1779-1791.

Zhang, B., Tian, J., Dong, D., Gu, D., Dong, Y., Zhang, L., Lian, Z., Liu, J., Luo, X., Pei, S., Mo, X., Huang, W., Ouyang, F., Guo, B., Liang, L., Chen, W., Liang, C. and Zhang, S. (2017) 'Radiomics Features of Multiparametric MRI as Novel Prognostic Factors in Advanced Nasopharyngeal Carcinoma', *Clinical Cancer Research,* 23(15), pp. 4259-4269.

Zhang, D. and Gu, M. (2023) 'Metabolic/endocrine disorders in survivors of childhood-onset and cranial radiotherapy- treated ALL/NHL: a meta-analysis', *Reproductive Biology and Endocrinology,* 21(1), pp. 91.

Zharmagambetov, A., Hada, S. S., Gabidolla, M. and Carreira-Perpiñán, M. Á. 'Non-Greedy Algorithms for Decision Tree Optimization: An Experimental Comparison'. *2021 International Joint Conference on Neural Networks (IJCNN)*, 18-22 July 2021, 1-8.

Zheng, Y. Z., Pan, L. L., Li, J., Chen, Z. S., Hua, X. L., Le, S. H., Zheng, H., Chen, C. and Hu, J. D. (2021) '[Clinical features and prognosis of ETV6-RUNX1-positive childhood B-precursor acute lymphocyte leukemia]', *Zhonghua xue ye xue za zhi = Zhonghua xueyexue zazhi,* 42(1), pp. 45-51.

Zhou, Z. H. (2012) *Ensemble Methods: Foundations and Algorithms. CHAPMAN & HALL/CRC MACHINE LEA*: Taylor & Francis.

Chapter 8. Supplementary

Supplementary Table 1. Mean dose intensity score, relative dose intensity score, and area under the curve dose intensity score for each major pathway.

Trial	Gender	Pathway	Total number of patients	DIS Mean	Relative DIS Mean	AUC DIS Mean
UKALLXI92	-	IT-MTX	310	2969.945	8.417126	65222.72
UKALLXI92	-	IT-MTX 3rd Block	305	3830.045	9.258546	67920.96
UKALLXI92	-	HD-MTX	379	6388.119	101.6981	90566.28
UKALLXI92	-	HD-MTX 3rd Block	388	7299.737	103.4396	93261.36
UKALL97	-	IT-MTX	149	3725.804	9.219065	58291.26
UKALL97	-	IT-MTX 3rd Block	522	4546.156	10.04801	62422.02
UKALL97	-	HD-MTX	22	7465.188	111.104	85301.08
UKALL97	-	HD-MTX 3rd Block	69	8158.428	108.6123	87709.61
UKALL97/99	Boys	Regimen A	327	4429.187	9.476793	93339.67
UKALL97/99	Girls	Regimen A	225	4503.288	9.522294	67656.29
UKALL97/99	Boys	Regimen B	128	5078.268	9.807615	100606.3
UKALL97/99	Girls	Regimen B	96	5030.447	9.776126	70127.16
UKALL97/99	Boys	Regimen C	81	4963.047	11.31385	117281.1
UKALL97/99	Girls	Regimen C	57	4967.822	11.25777	88078.03
UKALL2003	Boys	Regimen A 1 DI	271	3155.62	6.314819	106707.5
UKALL2003	Girls	Regimen A 1 DI	221	3160.693	6.327886	73334.58
UKALL2003	Boys	Regimen A 2 DI	570	4308.406	9.594157	107006.4
UKALL2003	Girls	Regimen A 2 DI	475	4306.945	9.582844	73308.11
UKALL2003	Boys	Regimen B 1 DI	150	3592.254	6.561718	110057.6
UKALL2003	Girls	Regimen B 1 DI	122	3598.669	6.571328	76680.39
UKALL2003	Boys	Regimen B 2 DI	359	4731.47	9.836128	110360.4
UKALL2003	Girls	Regimen B 2 DI	211	4728.394	9.808467	76655.17
UKALL2003	Boys	Regimen C 2 DI	350	3813.46	12.51463	111285.8
UKALL2003	Girls	Regimen C 2 DI	271	3816.092	12.57481	77606.94
UKALL2011	Boys	Regimen A SIM no pulses	51	3102.917	5.912716	95094.84

Trial	Gender	Pathway	Total number of patients	DIS Mean	Relative DIS Mean	AUC DIS Mean
UKALL2011	Girls	Regimen A SIM no pulses	32	3092.257	5.895517	65511.95
UKALL2011	Boys	Regimen A SIM pulses	176	3153.51	5.934528	106416.3
UKALL2011	Girls	Regimen A SIM pulses	179	3154.166	5.929325	72738.68
UKALL2011	Boys	Regimen A HDM no pulses	44	4951.68	86.43588	112554.9
UKALL2011	Girls	Regimen A HDM no pulses	41	4957.571	86.47898	83110.68
UKALL2011	Boys	Regimen A HDM pulses	45	5024.188	86.54027	124031.9
UKALL2011	Girls	Regimen A HDM pulses	30	5002.843	86.473	90198.65
UKALL2011	Boys	Regimen B SIM no pulses	32	3545.162	6.124535	99031.93
UKALL2011	Girls	Regimen B SIM no pulses	23	3549.468	6.138439	69574.96
UKALL2011	Boys	Regimen B SIM pulses	95	3593.299	6.141642	110331.5
UKALL2011	Girls	Regimen B SIM pulses	84	3596.104	6.148249	76665.22
UKALL2011	Boys	Regimen B HDM no pulses	34	5404.489	86.75048	116577.1
UKALL2011	Girls	Regimen B HDM no pulses	23	5397.78	86.65887	87023.38
UKALL2011	Boys	Regimen B HDM pulses	35	5457.283	86.75785	127917.2

Trial	Gender	Pathway	Total number of patients	DIS Mean	Relative DIS Mean	AUC DIS Mean
UKALL2011	Girls	Regimen B HDM pulses	31	5442.472	86.76857	94110.26
UKALL2011	Boys	Regimen C Capizzi no pulses	77	2867.592	7.497891	100179.2
UKALL2011	Girls	Regimen C Capizzi no pulses	57	2848.082	6.929655	70702.63
UKALL2011	Boys	Regimen C Capizzi pulses	194	2909.032	7.497681	111489
UKALL2011	Girls	Regimen C Capizzi pulses	146	2904.874	7.298534	77800.96
UKALL2011	Boys	Regimen C HDM no pulses	79	5468.094	97.17908	121610.1
UKALL2011	Girls	Regimen C HDM no pulses	44	5461.388	97.11375	92062.31
UKALL2011	Boys	Regimen C HDM pulses	78	5514.158	97.28083	132897.4
UKALL2011	Girls	Regimen C HDM pulses	56	5508.164	97.17373	99202.05

Supplementary Table 1. Mean dose intensity score, relative dose intensity score, and area under the curve dose intensity score for each major pathway. DIS: dose intensity score, AUC: area under the curve, IT-MTX: intrathecal methotrexate, HD-MTX: high dose methotrexate, DI: delayed intensification, SIM: standard interim maintenance, HDM: high dose maintenance.

Supplementary Table 2. Summary of the accuracy, F1-scores, ROC area under the curve, K-fold cross-validation minimum and maximum accuracy scores, the root features, 2nd split features, pruning technique, imbalanced class solution, and ranking of every decision tree produced in the *ETV6*::*RUNX1* subgroup.

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
Chestnut Features: Trial, Total Drug Dosages. Class: 4 class outcome										
1	0.89	0.97, 0.17, 0.04, 0.00	0.92	0.84	0.92	Cyt ≤ 2460	Vin ≤ 87.45 St ≤ 8229.575	None	None	Poor
2	0.92	0.96, 0.00, 0.08, 0.00	0.96	0.87	0.95	Cyt ≤ 2460	St ≤ 8829.575 -	Cost Complexity Pruning: alpha = 0.00175	None	Poor
3	0.92	1, 0.00, 0.00, 0.00	0.96	0.87	0.95	Cyt ≤ 2460	Cyt ≤ 1240.0 Asp ≤ 17500.0	GridSearch CV: Max depth = 2, Max features = 4	None	Poor
4	0.92	0.96, 0.00, 0.07, 0.00	0.95	0.86	0.95	Cyt ≤ 2460	Vin ≤ 87.45 St ≤ 8229.575	Max depth 5	None	Poor
5	0.53	0.68, 0.09, 0.12, 0.07	0.84	0.46	0.58	Asp ≤ 20000.0	Anth ≤ 233.75 St ≤ 6955.85	None	Balanced class weights	Poor
6	0.6	0.75, 0.06, 0.12, 0.03	0.73	0.22	0.49	Asp ≤ 20000.0	Anth ≤ 233.75 St ≤ 6955.85	Cost Complexity Pruning: alpha = 0.006	Balanced class weights	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
7	0.74	0.86, 0.00, 0.00, 0.02	0.75	0.21	0.36	Asp ≤ 20000.0	Cyt ≤ 2080.0 Meth ≤ 3133.2	GridSearch CV: Max depth = 2, Max features = 2	Balanced class weights	Poor
8	0.36	0.55, 0.03, 0.00, 0.01	0.64	0.11	0.32	Asp ≤ 20000.0	Anth ≤ 233.75 -	Max depth 5	Balanced class weights	Poor
Oak Features: Trial, Reg, DI, steroid received and purine received. Class: 4 class outcome										
9	0.9	0.95, 0.00, 0.00, 0.00	0.94	0.88	0.94	Reg ≤ 1.5	Purine ≤ 0.5 Trial ≤ 2.5	None	None	Poor
10	0.9	0.95, 0.00, 0.00, 0.00	0.95	0.88	0.94	Reg ≤ 1.5	- Trial ≤ 2.5	Cost Complexity Pruning: alpha = 0.0007	None	Poor
11	0.9	0.95, 0.00, 0.00, 0.00	0.95	0.88	0.94	Steroid ≤ 0.5	DI ≤ 1.5 Purine ≤ 0.5	GridSearch CV: Max depth = 2, Max features = 1	None	Poor
12	0.9	0.95, 0.00, 0.00, 0.00	0.95	0.88	0.94	Reg ≤ 1.5	Purine ≤ 0.5 DI ≤ 1.5	Max depth 5	None	Poor
13	0.57	0.74, 0.03, 0.00, 0.11	0.68	0.12	0.37	DI ≤ 1.5	Reg ≤ 0.5 Purine ≤ 0.5	None	Balanced class weights	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
14	0.57	0.74, 0.03, 0.00, 0.12	0.65	0.01	0.76	DI ≤ 1.5	Reg ≤ 0.5 Purine ≤ 0.5	Cost complexity pruning: alpha = 0.01	Balanced class weights	Poor
15	0.59	0.75, 0.06, 0.09, 0.13	0.65	0.15	0.73	DI ≤ 1.5	Reg ≤ 0.5 Reg ≤ 0.5	GridSearch CV: max depth = 3, max features = 3	Balanced class weights	Poor
16	0.56	0.73, 0.02, 0.00, 0.13	0.64	0.12	0.53	DI ≤ 1.5	Reg ≤ 0.5 Purine ≤ 0.5	Max depth 5	Balanced class weights	Poor
Elm Features: Trial, DI, steroid received and purine received. Class: 4 class outcome										
17	0.92	0.96, 0.00, 0.00, 0.00	0.96	0.87	0.95	DI ≤ 2.5	Steroid ≤ 0.5 Steorid ≤ 0.5	None	None	Poor
18	0.92	0.96, 0.00, 0.00, 0.00	0.96	0.87	0.95	-	-	Cost Complexity Pruning: alpha = 0.2	None	Poor
19	0.92	0.96, 0.00, 0.00, 0.00	0.96	0.87	0.95	DI ≤ 2.5	Steroid ≤ 0.5 Steroid ≤ 0.5	GridSearch CV: max depth = 2, max features = 1	None	Poor
20	0.92	0.96, 0.00, 0.00, 0.00	0.96	0.87	0.95	DI ≤ 2.5	Steroid ≤ 0.5 Steroid ≤ 0.5	Max depth 5	None	Poor
21	0.25	0.40, 0.01, 0.10, 0.03	0.59	0.18	0.29	Purine ≤ 0.5	Trial ≤ 1.5 DI ≤ 2.5	None	Balanced class weights	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
22	0.02	0.00, 0.00, 0.18, 0.02	0.5	0.05	0.79	Purine ≤ 0.5	Trial ≤ 1.5 -	Cost complexity pruning: alpha = 0.007	Balanced class weights	Poor
23	0.8	0.89, 0.17, 0.18, 0.00	0.78	0.39	0.76	Purine ≤ 0.5	Trial ≤ 1.5 DI ≤ 2.5	GridSearch CV: max depth = 2, max features = 3	Balanced class weights	Poor
24	0.25	0.40, 0.01, 0.10, 0.03	0.59	0.18	0.29	Purine ≤ 0.5	Trial ≤ 1.5 DI ≤ 2.5	Max depth 5	Balanced class weights	Poor
Chestnut Features: Trial, Total Drug Dosages. Class: 2 class outcome - RR										
25	0.92	0.96, 0.10	0.43	0.9	0.9	Cyt ≤ 2370.0	Asp ≤ 6500.0 St ≤ 7741.438	None	None	Poor
26	0.93	0.97, 0.06	0.55	0.9	0.9	Cyt ≤ 2370.0	- St ≤ 7741.438	Cost complexity pruning: alpha = 0.0016	None	Poor
27	0.93	0.96, 0.00	0.64	0.9	0.9	Asp ≤ 4250.0	Vin ≤ 68.25 Cyt ≤ 2460.0	GridSearch CV: max depth = 2, max features = 1	None	Poor
28	0.93	0.96, 0.00	0.61	0.9	1.0	Cyt ≤ 2370.0	Asp ≤ 6500.0 St ≤ 7741.438	Max depth 5	None	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
29	0.66	0.79, 0.10	0.41	0.7	0.8	Asp ≤ 4250.0	Meth ≤ 21972.5 Meth ≤ 2013.6	None	Balanced class weights	Poor
30	0.79	0.88, 0.22	0.63	0.6	0.8	Asp ≤ 4250.0	Meth ≤ 21972.5 Meth ≤ 2013.6	Cost complexity pruning: alpha = 0.0025	Balanced class weights	Poor
31	0.76	0.86, 0.12	0.46	0.6	0.7	Asp ≤ 4250.0	Meth ≤ 21972.5 Vin ≤ 24.75	GridSearch CV: max depth = 8, max features = 3	Balanced class weights	Poor
32	0.72	0.83, 0.19	0.71	0.5	0.8	Asp ≤ 4250.0	Meth ≤ 21972.5 Meth ≤ 2013.6	Max depth 5	Balanced class weights	Poor
Chestnut Features: Trial, Total Drug Dosages. Class: 4 class outcome										
33	0.83	0.79, 0.91, 0.76, 0.86	0.92	0.81	0.88	Cyc ≤ 1200.5	Purine ≤ 40586.0 Vin ≤ 68.5	None	Oversampling: SMOTE with auto sampling strategy	Poor
34	0.7	0.64, 0.80, 0.62, 0.74	0.91	0.67	0.74	Cyc ≤ 1200.5	Purine ≤ 40586.0 Vin ≤ 68.5	Cost complexity pruning: alpha = 0.003	Oversampling: SMOTE with auto sampling strategy	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
35	0.83	0.80, 0.89, 0.77, 0.85	0.93	0.81	0.87	Cyc ≤ 1200.5	Cyt ≤ 600.5 Anth ≤ 218.5	GridSearch CV: max depth = 14, max features = 4	Oversampling: SMOTE with auto sampling strategy	Poor
36	0.61	0.53, 0.72, 0.49, 0.69	0.85	0.62	0.65	Cyc ≤ 1200.5	Purine ≤ 40586.0 Vin ≤ 68.5	Max depth 5	Oversampling: SMOTE with auto sampling strategy	Poor
Chestnut Features: Trial, Total Drug Dosages. Class: 2 class outcome - RD										
37	0.98	0.99, 0.36	0.64	0.97	1.00	Asp ≤ 20000.0	St ≤ 2231.938 St ≤ 6955.85	None	None	Poor
38	0.98	0.99, 0.00	0.47	0.97	1.00	Asp ≤ 20000.0	St ≤ 2231.938 -	Cost complexity pruning: alpha = 0.0008	None	Poor
39	0.99	0.99, 0.00	0.46	0.97	1.00	Asp ≤ 20000.0	Cyt ≤ 1710.0 Meth ≤ 2460.0	GridSearch CV: max depth = 2, max features = 4	None	Poor
40	0.99	0.99, 0.00	0.5	0.97	1.00	St ≤ 2231.938	Asp ≤ 7000.0 Cyt ≤ 2460.0	Max depth 5	None	Poor
41	0.91	0.95, 0.09	0.64	0.82	0.91	Purine ≤ 78592.5	Cyt ≤ 1710.0 -	None	Balanced class weights	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
42	0.76	0.86, 0.04	0.57	0.12	0.46	Purine ≤ 78592.5	Cyt ≤ 1710.0 -	Cost complexity pruning: alpha = 0.01	Balanced class weights	Poor
43	0.91	0.95, 0.10	0.64	0.82	0.92	Asp ≤ 7700.0	Vin ≤ 49.875 St ≤ 9345.438	GridSearch CV: max depth = 13, max features = 1	Balanced class weights	Poor
44	0.65	0.79, 0.03	0.51	0.47	0.81	Purine ≤ 78592.5	Cyt ≤ 1710.0 -	Max depth 5	Balanced class weights	Poor
Chestnut Features: Trial, Total Drug Dosages. Class: 4 class outcome										
45	0.89	0.94, 0.86, 0.90, 0.85	0.93	0.85	0.92	Asp ≤ 3005.0	Cyc ≤ 1074.0 Purine ≤ 51231.0	None	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Good

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
46	0.85	0.93, 0.79, 0.86, 0.84	0.96	0.80	0.88	Asp ≤ 3005.0	- Purine ≤ 51231.0	Cost complexity pruning: alpha = 0.01	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Poor
47	0.87	0.94, 0.83, 0.88, 0.84	0.96	0.82	0.92	Asp ≤ 3005.0	Cyt ≤ 644.5 Purine ≤ 51231.0	GridSearch CV: max depth = 7, max features = 7	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Good

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
48	0.86	0.93, 0.81, 0.86, 0.84	0.97	0.80	0.90	Asp ≤ 3005.0	Cyc ≤ 1074.0 Purine ≤ 51231.0	Max depth 5	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Good
49	0.52	0.75, 0.36, 0.50, 0.53	0.66	0.38	0.71	Meth ≤ 2048.5	Meth ≤ 2037.5 Meth ≤ 2160.0	None	Undersampling: Nearmiss version 1	Poor
50	0.52	0.75, 0.46, 0.46, 0.50	0.78	0.25	0.75	Meth ≤ 2048.5	Meth ≤ 2037.5 Meth ≤ 2160.0	Cost complexity pruning: alpha = 0.02	Undersampling: Nearmiss version 1	Poor
51	0.52	0.67, 0.36, 0.50, 0.57	0.66	0.38	1.00	Asp ≤ 3400.0	Trial ≤ 3.5 Meth ≤ 2120.5	GridSearch CV: max depth = 4, max features = 3	Undersampling: Nearmiss version 1	Poor
52	0.26	0.33, 0.43, 0.00, 0.00	0.55	0.13	0.75	Meth ≤ 2048.2	-	Max depth 5	Undersampling: Nearmiss version 1	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
Oak Features: Trial, Reg, DI, steroid received, purine received. Class: 2 class outcome RR										
53	0.91	0.95, 0.00	0.51	0.91	0.95	Reg ≤ 1.5	Steroid ≤ 0.5 Steroid ≤ 0.5	None	None	Poor
54	0.91	0.95, 0.00	0.51	0.89	0.95	Reg ≤ 1.5	-	Cost complexity pruning: alpha = 0.001	None	Poor
55	0.91	0.95, 0.00	0.55	0.89	0.95	Steroid ≤ 0.5	DI ≤ 1.5 Purine ≤ 0.5	GridSearch CV: max depth = 2, max features = 1	None	Poor
56	0.91	0.95, 0.00	0.54	0.89	0.95	Reg ≤ 1.5	Steroid ≤ 0.5 Trial ≤ 3.5	Max depth 5	None	Poor
57	0.73	0.84, 0.17	0.51	0.57	0.70	Reg ≤ 1.5	Steroid ≤ 0.5 Steroid ≤ 0.5	None	Balanced class weights	Poor
58	0.81	0.89, 0.12	0.51	0.60	0.87	Reg ≤ 1.5	-	Cost complexity pruning: alpha = 0.006	Balanced class weights	Poor
59	0.76	0.86, 0.16	0.54	0.55	0.81	Reg ≤ 1.5	Steroid ≤ 0.5 Steroid ≤ 0.5	GridSearch CV: max depth = 2, max features = 4	Balanced class weights	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
60	0.73	0.84, 0.18	0.54	0.57	0.69	Reg ≤ 1.5	Steroid ≤ 0.5 Trial ≤ 3.5	Max depth 5	Balanced class weights	Poor
Oak Features: Trial, Reg, DI, steroid received, purine received. Class: 4 class outcome										
61	0.42	0.23, 0.45, 0.45, 0.51	0.73	0.40	0.44	DI ≤ 1.5	Reg ≤ 0.5 Reg ≤ 0.5	None	Oversampling: SMOTE with auto sampling strategy	Poor
62	0.42	0.23, 0.43, 0.45, 0.51	0.71	0.40	0.43	DI ≤ 1.5	Reg ≤ 0.5 Reg ≤ 0.5	Cost complexity pruning: alpha = 0.0025	Oversampling: SMOTE with auto sampling strategy	Poor
63	0.42	0.23, 0.45, 0.45, 0.51	0.73	0.40	0.43	DI ≤ 1.5	Reg ≤ 0.5 Reg ≤ 0.5	GridSearch CV: max depth = 5, max features = 3	Oversampling: SMOTE with auto sampling strategy	Poor
64	0.42	0.23, 0.45, 0.45, 0.51	0.73	0.40	0.43	DI ≤ 1.5	Reg ≤ 0.5 Reg ≤ 0.5	Max depth 5	Oversampling: SMOTE with auto sampling strategy	Poor
Oak Features: Trial, Reg, DI, steroid received, purine received. Class: 2 class outcome - RD										
65	0.98	0.99, 0.00	0.65	0.97	1.00	Reg ≤ 1.5	Purine ≤ 0.5 Trial ≤ 2.5	None	None	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
66	0.98	0.99, 0.00	0.5	0.97	1.00	-	-	Cost complexity pruning: alpha = 0.0008	None	Poor
67	0.98	0.99, 0.00	0.3	0.97	1.00	Steroid ≤ 0.5	DI ≤ 1.5 Purine ≤ 0.5	GridSearch CV: max depth = 2, max features = 1	None	Poor
68	0.98	0.99, 0.00	0.64	0.97	1.00	Reg ≤ 1.5	Purine ≤ 0.5 Trial ≤ 3.5	Max depth 5	None	Poor
69	0.81	0.89, 0.03	0.65	0.52	0.65	Reg ≤ 0.5	Trial ≤ 3.5 DI ≤ 1.5	None	Balanced class weights	Poor
70	0.82	0.90, 0.05	0.7	0.45	0.92	Reg ≤ 0.5	Trial ≤ 3.5 DI ≤ 1.5	Cost complexity pruning: alpha = 0.018	Balanced class weights	Poor
71	0.8	0.89, 0.05	0.78	0.50	0.84	Reg ≤ 0.5	Trial ≤ 3.5 DI ≤ 1.5	GridSearch CV: max depth = 5, max features = 3	Balanced class weights	Poor
72	0.76	0.86, 0.04	0.74	0.43	0.57	Reg ≤ 0.5	Trial ≤ 3.5 DI ≤ 1.5	Max depth 5	Balanced class weights	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
Oak Features: Trial, Reg, DI, steroid received, purine received. Class: 4 class outcome										
73	0.44	0.51, 0.55, 0.00, 0.46	0.72	0.39	0.51	DI ≤ 1.5	Trial ≤ 3.5 Trial ≤ 2.5	None	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Poor
74	0.4	0.51, 0.32, 0.00, 0.45	0.69	0.34	0.46	DI ≤ 1.5	- Trial ≤ 2.5	Cost complexity pruning: alpha = 0.034	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
75	0.42	0.51, 0.56, 0.00, 0.33	0.71	0.37	0.51	DI ≤ 1.5	Reg ≤ 1.5 Reg ≤ 0.5	GridSearch CV: max depth = 3, max features = 2	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Poor
76	0.42	0.51, 0.47, 0.00, 0.44	0.7	0.39	0.51	DI ≤ 1.5	Trial ≤ 3.5 Trial ≤ 2.5	Max depth 5	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Poor
77	0.28	0.43, 0.40, 0.00, 0.20	0.56	0.17	0.67	DI ≤ 1.5	Reg ≤ 0.5 Trial ≤ 2.5	None	Undersampling: Nearmiss version 1	Poor
78	0.28	0.43, 0.40, 0.00, 0.20	0.56	0.17	0.67	DI ≤ 1.5	Reg ≤ 0.5 Trial ≤ 2.5	Cost complexity pruning: alpha = 0.019	Undersampling: Nearmiss version 1	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
79	0.28	0.43, 0.40, 0.00, 0.20	0.56	0.17	0.67	DI ≤ 1.5	Reg ≤ 0.5 Trial ≤ 2.5	GridSearch CV: max depth = 3, max features = 5	Undersampling: Nearmiss version 1	Poor
80	0.28	0.43, 0.40, 0.00, 0.20	0.56	0.00	0.67	DI ≤ 1.5	Reg ≤ 0.5 Trial ≤ 2.5	Max depth 5	Undersampling: Nearmiss version 1	Poor
Elm Features: Trial, DI, steroid received, purine received. Class: 2 class outcome - RR										
81	0.93	0.96, 0.00	0.51	0.89	0.96	Steroid ≤ 0.5	Trial ≤ 3.5 DI ≤ 2.5	None	None	Poor
82	0.93	0.96, 0.00	0.55	0.89	0.96	Steroid ≤ 0.5	-	Cost complexity pruning: alpha = 0.0016	None	Poor
83	0.93	0.96, 0.00	0.57	0.89	0.96	DI ≤ 2.5	Steroid ≤ 0.5 Steroid ≤ 0.5	GridSearch CV: max depth = 2, max features = 1	None	Poor
84	0.93	0.96, 0.00	0.55	0.89	0.96	Steroid ≤ 0.5	Trial ≤ 3.5 DI ≤ 2.5	Max depth 5	None	Poor
85	0.8	0.89, 0.12	0.51	0.58	0.81	Steroid ≤ 0.5	Trial ≤ 3.5 DI ≤ 2.5	None	Balanced class weights	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
86	0.83	0.90, 0.16	0.55	0.77	0.86	Steroid ≤ 0.5	- DI ≤ 2.5	Cost complexity pruning: alpha = 0.004	Balanced class weights	Poor
87	0.83	0.90, 0.16	0.58	0.58	0.83	Steroid ≤ 0.5	Trial ≤ 3.5 DI ≤ 2.5	GridSearch CV: max depth = 2, max features = 4	Balanced class weights	Poor
88	0.83	0.91, 0.12	0.55	0.58	0.82	Steroid ≤ 0.5	Trial ≤ 3.5 DI ≤ 2.5	Max depth 5	Balanced class weights	Poor
Elm Features: Trial, DI, steroid received, purine received. Class: 4 class outcome										
89	0.39	0.04, 0.46, 0.43, 0.42	0.67	0.37	0.43	DI ≤ 2.5	DI ≤ 1.5 Purine ≤ 0.5	None	Oversampling: SMOTE with auto sampling strategy	Poor
90	0.39	0.00, 0.46, 0.43, 0.44	0.67	0.37	0.43	DI ≤ 2.5	DI ≤ 1.5 Purine ≤ 0.5	Cost complexity pruning: alpha = 0.002	Oversampling: SMOTE with auto sampling strategy	Poor
91	0.39	0.04, 0.46, 0.43, 0.42	0.67	0.37	0.43	DI ≤ 2.5	Steroid ≤ 0.5 Steroid ≤ 0.5	GridSearch CV: max depth = 5, max features = 1	Oversampling: SMOTE with auto sampling strategy	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
92	0.37	0.04, 0.41, 0.43, 0.42	0.66	0.37	0.43	DI ≤ 2.5	DI ≤ 1.5 Purine ≤ 0.5	Max depth 5	Oversampling: SMOTE with auto sampling strategy	Poor
Elm Features: Trial, DI, steroid received, purine received. Class: 2 class outcome - RD										
93	0.99	0.99, 0.00	0.49	0.97	1.00	Purine ≤ 0.5	Trial ≤ 3.5 DI ≤ 2.5	None	None	Poor
94	0.99	0.99, 0.00	0.5	0.97	1.00	-	-	Cost complexity pruning: alpha = 0.0008	None	Poor
95	0.99	0.99, 0.00	0.58	0.97	1.00	DI ≤ 2.5	Steroid ≤ 0.5 Steroid ≤ 0.5	GridSearch CV: max depth = 2, max features = 1	None	Poor
96	0.99	0.99, 0.00	0.48	0.97	1.00	Purine ≤ 0.5	Trial ≤ 3.5 DI ≤ 2.5	Max depth 5	None	Poor
97	0.59	0.74, 0.02	0.49	0.42	0.90	Purine ≤ 0.5	Trial ≤ 1.5 DI ≤ 2.5	None	Balanced class weights	Poor
98	0.91	0.95, 0.09	0.63	0.02	0.94	Purine ≤ 0.5	-	Cost complexity pruning: alpha = 0.012	Balanced class weights	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
99	0.87	0.93, 0.07	0.64	0.32	0.95	DI ≤ 2.5	Steroid ≤ 0.5 Steroid ≤ 0.5	GridSearch CV: max depth = 3, max features = 1	Balanced class weights	Poor
100	0.58	0.74, 0.02	0.48	0.41	0.90	Purine ≤ 0.5	Trial ≤ 1.5 DI ≤ 2.5	Max depth 5	Balanced class weights	Poor
Elm Features: Trial, DI, steroid received, purine received. Class: 4 class outcome										
101	0.4	0.46, 0.71, 0.00, 0.00	0.71	0.34	0.52	Trial ≤ 2.5	- DI ≤ 1.5	None	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Poor
102	0.4	0.46, 0.71, 0.00, 0.00	0.71	0.34	0.52	Trial ≤ 2.5	- DI ≤ 1.5	Cost complexity pruning: alpha = 0.04	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
103	0.4	0.46, 0.71, 0.00, 0.00	0.71	0.34	0.52	DI ≤ 1.5	Trial ≤ 3.5 Trial ≤ 2.5	GridSearch CV: max depth = 2, max features = 1	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Poor
104	0.4	0.46, 0.71, 0.00, 0.00	0.71	0.34	0.52	Trial ≤ 2.5	- DI ≤ 1.5	Max depth 5	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Poor
105	0.35	0.32, 0.77, 0.00, 0.00	0.72	0.14	0.43	Trial ≤ 2.5	- DI ≤ 1.5	None	Undersampling: Nearmiss version 1	Poor
106	0.42	0.32, 0.77, 0.00, 0.00	0.72	0.14	0.43	Trial ≤ 2.5	- DI ≤ 1.5	Cost complexity pruning: alpha = 0.035	Undersampling: Nearmiss version 1	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
107	0.35	0.32, 0.77, 0.00, 0.00	0.72	0.00	0.43	DI ≤ 1.5	Trial ≤ 3.5 Trial ≤ 2.5	GridSearch CV: max depth = 3, max features = 1	Undersampling: Nearmiss version 1	Poor
108	0.35	0.32, 0.77, 0.00, 0.00	0.72	0.14	0.43	Trial ≤ 2.5	- DI ≤ 1.5	Max depth 5	Undersampling: Nearmiss version 1	Poor
Chestnut Features: Total drug dosages, Trial. Class: 2 Class outcome RR										
109	0.95	0.95, 0.95	0.97	0.92	1.00	St ≤ 9621.5	- Meth ≤ 3087.0	None	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Good

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
110	0.95	0.95, 0.95	0.95	0.92	0.99	St ≤ 9621.5	- Meth ≤ 3087.0	Cost complexity pruning: alpha = 0.02	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Good
111	0.94	0.94, 0.94	0.95	0.93	1.00	Asp ≤ 4036.5	Vin ≤ 67.5 -	GridSearch CV: max depth = 8, max features = 1	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Good

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
112	0.95	0.95, 0.95	0.97	0.92	0.99	St ≤ 9621.5	- Meth ≤ 3087.0	Max depth 5	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Good
113	0.99	0.99, 0.99	0.98	0.84	1.00	Purine ≤ 77894.0	- Steroid ≤ 9556.0	None	Undersampling: Nearmiss version 1	Good
114	0.99	0.99, 0.99	0.99	0.84	1.00	Purine ≤ 77894.0	- St ≤ 9556.0	Cost complexity pruning: alpha = 0.015	Undersampling: Nearmiss version 1	Good
115	0.99	0.99, 0.99	0.98	0.84	1.00	Purine ≤ 77894.0	- Cyt ≤ 1410.0	GridSearch CV: max depth = 6, max features = 4	Undersampling: Nearmiss version 1	Good
116	0.89	0.91, 0.88	0.96	0.68	1.00	Purine ≤ 77894.0	- Purine ≤ 78312.5	Max depth 5	Undersampling: Nearmiss version 1	Good

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
117	0.85	0.87, 0.84	0.9	0.79	0.86	Cyt ≤ 1271.5	Purine ≤ 53833.5 Anth ≤ 249.5	None	Oversampling: SMOTE with auto sampling strategy	Good
118	0.81	0.82, 0.79	0.87	0.73	0.86	Cyt ≤ 1271.5	Purine ≤ 53833.5 Anth ≤ 249.5	Cost complexity pruning: alpha = 0.002	Oversampling: SMOTE with auto sampling strategy	Poor
119	0.83	0.85, 0.81	0.89	0.79	0.86	Cyt ≤ 1271.5	Purine ≤ 53833.5 Cyc ≤ 2996.5	GridSearch CV: max depth = 14, max features = 6	Oversampling: SMOTE with auto sampling strategy	Good
120	0.71	0.73, 0.69	0.78	0.65	0.76	Cyt ≤ 1271.5	Purine ≤ 53833.5 Anth ≤ 249.5	Max depth 5	Oversampling: SMOTE with auto sampling strategy	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
Chestnut Features: Total drug dosages, Trial. Class: 2 Class outcome RD										
121	0.98	0.98, 0.98	0.98	0.95	0.99	Meth ≤ 2048.5	Cyt ≤ 1205.5 -	None	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Good
122	0.98	0.98, 0.98	0.98	0.95	0.99	Meth ≤ 2048.5	Cyt ≤ 1205.5 -	Cost complexity pruning: alpha = 0.015	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Good

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
123	0.98	0.98, 0.98	0.98	0.95	0.99	Cyt ≤ 1201.5	Purine ≤ 50871.0 -	GridSearch CV: max depth = 2, max features = 3	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Good
124	0.98	0.98, 0.98	0.98	0.95	0.99	Meth ≤ 2048.5	Cyt ≤ 1205.5 -	Max depth 5	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Good
125	0.67	0.60, 0.71	0.78	0.50	1.00	Cyt ≤ 1500.0	Meth ≤ 2515.0 -	None	Undersampling: Nearmiss version 1	Poor
126	0.67	0.60, 0.71	0.78	0.50	1.00	Cyt ≤ 1500.0	Meth ≤ 2515.0 -	Cost complexity pruning: alpha = 0.1	Undersampling: Nearmiss version 1	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
127	0.99	0.99, 0.99	0.72	0.25	1.00	Cyt ≤ 1500.0	Purine ≤ 64123.5 -	GridSearch CV: max depth = 2, max features = 3	Undersampling: Nearmiss version 1	Poor
128	0.25	0.40, 0.00	0.5	0.00	0.50	-	-	Max depth 5	Undersampling: Nearmiss version 1	Poor
129	0.97	0.97, 0.97	0.99	0.95	0.98	Cyc ≤ 1001.0	- Vin ≤ 67.5	None	Oversampling: SMOTE with auto sampling strategy	Good
130	0.93	0.93, 0.93	0.98	0.94	0.97	Cyc ≤ 1001.0	- Vin ≤ 67.5	Cost complexity pruning: alpha = 0.005	Oversampling: SMOTE with auto sampling strategy	Good
131	0.96	0.96, 0.96	0.98	0.95	0.99	Cyc ≤ 1001.0	- St ≤ 10333.0	GridSearch CV: max depth = 13, max features = 3	Oversampling: SMOTE with auto sampling strategy	Good
132	0.89	0.90, 0.89	0.95	0.86	0.90	Cyc ≤ 1001.0	- Vin ≤ 67.5	Max depth 5	Oversampling: SMOTE with auto sampling strategy	Good

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
Elm Features: Trial, DIs, St, Purine. Class: 2 Class outcome RR										
133	0.78	0.80, 0.76	0.81	0.72	0.85	Trial ≤ 3.5	Trial ≤ 2.5 -	None	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Poor
134	0.78	0.81, 0.73	0.79	0.71	0.83	Trial ≤ 3.5	Trial ≤ 2.5 -	Cost complexity pruning: alpha = 0.01	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
135	0.78	0.80, 0.76	0.82	0.72	0.85	DI ≤ 1.5	Trial ≤ 3.5 Trial ≤ 2.5	GridSearch CV: max depth = 3, max features = 4	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Poor
136	0.78	0.80, 0.76	0.81	0.72	0.85	Trial ≤ 3.5	Trial ≤ 2.5 -	Max depth 5	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Poor
137	0.68	0.67, 0.70	0.6	0.60	0.95	Trial ≤ 3.5	Trial ≤ 2.5 -	None	Undersampling: Nearmiss version 1	Poor
138	0.8	0.83, 0.75	0.8	0.60	1.00	Trial ≤ 3.5	Trial ≤ 2.5 -	Cost complexity pruning: alpha = 0.05	Undersampling: Nearmiss version 1	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
139	0.68	0.67, 0.70	0.8	0.50	0.95	Trial ≤ 3.5	Trial ≤ 2.5 -	GridSearch CV: max depth = 4, max features = 5	Undersampling: Nearmiss version 1	Poor
140	0.52	0.49, 0.54	0.61	0.50	0.85	Trial ≤ 3.5	Reg ≤ 0.5 -	Max depth 5	Undersampling: Nearmiss version 1	Poor
141	0.61	0.65, 0.56	0.62	0.58	0.67	Reg ≤ 1.5	Trial ≤ 2.5 Trial ≤ 2.5	None	Oversampling: SMOTE with auto sampling strategy	Poor
142	0.57	0.62, 0.5	0.58	0.55	0.63	Reg ≤ 1.5	Trial ≤ 2.5 -	Cost complexity pruning: alpha = 0.0045	Oversampling: SMOTE with auto sampling strategy	Poor
143	0.61	0.65, 0.56	0.62	0.58	0.67	Steroid ≤ 0.5	DI ≤ 1.5 Purine ≤ 0.5	GridSearch CV: max depth = 5, max features = 1	Oversampling: SMOTE with auto sampling strategy	Poor
144	0.73	0.65, 0.55	0.61	0.58	0.67	Reg ≤ 1.5	Trial ≤ 2.5 Trial ≤ 2.5	Max depth 5	Oversampling: SMOTE with auto sampling strategy	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
Elm Features: Trial, DIs, St, Purine. Class: 2 Class outcome RD										
145	0.8	0.83, 0.76	0.81	0.77	0.87	DI ≤ 1.5	- Trial ≤ 2.5	None	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Poor
146	0.8	0.83, 0.76	0.81	0.77	0.87	DI ≤ 1.5	- Trial ≤ 2.5	Cost complexity pruning: alpha = 0.03	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
147	0.8	0.83, 0.76	0.81	0.77	0.87	DI ≤ 1.5	- Trial ≤ 2.5	GridSearch CV: max depth = 2, max features = 3	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Poor
148	0.8	0.83, 0.76	0.81	0.77	0.87	DI ≤ 1.5	- Trial ≤ 2.5	Max depth 5	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Poor
149	0.56	0.60, 0.50	0.67	0.33	1.00	Reg ≤ 0.5	Steroid ≤ 0.5 -	None	Undersampling: Nearmiss version 1	Poor
150	0.56	0.60, 0.50	0.67	0.33	1.00	Reg ≤ 0.5	Steroid ≤ 0.5 -	Cost complexity pruning: alpha = 0.04	Undersampling: Nearmiss version 1	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
151	0.67	0.67, 0.67	0.75	0.33	1.00	Steroid ≤ 0.5	DI ≤ 1.5 -	GridSearch CV: max depth = 2, max features = 1	Undersampling: Nearmiss version 1	Poor
152	0.33	0.50, 0.00	0.5	0.00	0.33	-	-	Max depth 5	Undersampling: Nearmiss version 1	Poor
153	0.71	0.68, 0.74	0.77	0.66	0.77	Trial ≤ 3.5	Purine ≤ 0.5 Reg ≤ 0.5	None	Oversampling: SMOTE with auto sampling strategy	Poor
154	0.68	0.54, 0.75	0.74	0.66	0.74	Trial ≤ 3.5	Purine ≤ 0.5 Reg ≤ 0.5	Cost complexity pruning: alpha = 0.005	Oversampling: SMOTE with auto sampling strategy	Poor
155	0.71	0.68, 0.74	0.77	0.66	0.77	Steroid ≤ 0.5	DI ≤ 1.5 Purine ≤ 0.5	GridSearch CV: max depth = 5, max features = 1	Oversampling: SMOTE with auto sampling strategy	Poor
156	0.63	0.48, 0.71	0.69	0.60	0.70	Trial ≤ 3.5	Purine ≤ 0.5 Reg ≤ 0.5	Max depth 5	Oversampling: SMOTE with auto sampling strategy	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
Oak Features: Trial, Reg, DIs, St, Purine. Class: 2 Class outcome RR										
157	0.81	0.84, 0.77	0.86	0.72	0.85	Trial ≤ 2.5	- Trial ≤ 3.5	None	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Poor
158	0.81	0.84, 0.77	0.81	0.72	0.85	Trial ≤ 2.5	- Trial ≤ 3.5	Cost complexity pruning: alpha = 0.01	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
159	0.81	0.84, 0.77	0.81	0.72	0.85	Trial ≤ 2.5	- Trial ≤ 3.5	GridSearch CV: max depth = 2, max features = 4	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Poor
160	0.81	0.84, 0.77	0.86	0.72	0.85	Trial ≤ 2.5	- Trial ≤ 3.5	Max depth 5	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Poor
161	0.8	0.80, 0.81	0.88	0.64	0.85	Trial ≤ 2.5	- Trial ≤ 3.5	None	Undersampling: Nearmiss version 1	Good
162	0.86	0.88, 0.83	0.85	0.56	0.96	Trial ≤ 2.5	- Trial ≤ 3.5	Cost complexity pruning: alpha = 0.04	Undersampling: Nearmiss version 1	Good

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
163	0.8	0.80, 0.81	0.88	0.56	0.92	Trial ≤ 2.5	- Trial ≤ 3.5	GridSearch CV: max depth = 3, max features = 4	Undersampling: Nearmiss version 1	Good
164	0.8	0.80, 0.81	0.88	0.56	0.92	Trial ≤ 2.5	- Trial ≤ 3.5	Max depth 5	Undersampling: Nearmiss version 1	Good
165	0.59	0.66, 0.47	0.59	0.51	0.63	Steroid ≤ 0.5	Purine ≤ 0.5 DI ≤ 2.5	None	Oversampling: SMOTE with auto sampling strategy	Poor
166	0.58	0.65, 0.48	0.6	0.51	0.62	Steroid ≤ 0.5	Purine ≤ 0.5 DI ≤ 2.5	Cost complexity pruning: alpha = 0.001	Oversampling: SMOTE with auto sampling strategy	Poor
167	0.56	0.66, 0.40	0.57	0.51	0.61	Steroid ≤ 0.5	Trial ≤ 1.5 Trial ≤ 1.5	GridSearch CV: max depth = 3, max features = 2	Oversampling: SMOTE with auto sampling strategy	Poor
168	0.59	0.67, 0.47	0.59	0.51	0.63	Steroid ≤ 0.5	Purine ≤ 0.5 DI ≤ 2.5	Max depth 5	Oversampling: SMOTE with auto sampling strategy	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
Oak Features: Trial, Reg, DIs, St, Purine. Class: 2 Class outcome RD										
169	0.86	0.88, 0.84	0.86	0.80	0.90	DI ≤ 1.5	- Trial ≤ 2.5	None	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Good
170	0.86	0.88, 0.84	0.86	0.80	0.90	DI ≤ 1.5	- Trial ≤ 2.5	Cost complexity pruning: alpha = 0	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Good

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
171	0.86	0.88, 0.84	0.86	0.80	0.90	DI ≤ 1.5	- Trial ≤ 2.5	GridSearch CV: max depth = 2, max features = 2	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Good
172	0.86	0.88, 0.84	0.86	0.80	0.90	DI ≤ 1.5	- Trial ≤ 2.5	Max depth 5	Undersampling and Oversampling: NearMiss version 1 and a ratio sampling strategy of 486 patients in each class	Good
173	0.67	0.60, 0.71	0.78	0.50	1.00	Trial ≤ 2.5	- DI ≤ 1.5	None	Undersampling: Nearmiss version 1	Poor
174	0.67	0.60, 0.71	0.78	0.50	1.00	Trial ≤ 2.5	- DI ≤ 1.5	Cost complexity pruning: alpha = 0.04	Undersampling: Nearmiss version 1	Poor

Tree	Acc	F1-Score	AUC	CV Min	CV Max	Root Feature	2nd Split Features	Pruning	Imbalanced Class solution	Ranking
175	0.67	0.60, 0.71	0.78	0.50	1.00	DI ≤ 2.5	DI ≤ 1.5 -	GridSearch CV: max depth = 3, max features = 3	Undersampling: Nearmiss version 1	Poor
176	0.25	0.40, 0.00	0.5	0.00	0.50	-	-	Max depth 5	Undersampling: Nearmiss version 1	Poor
177	0.61	0.65, 0.56	0.68	0.59	0.69	Trial ≤ 3.5	Trial ≤ 1.5 -	None	Oversampling: SMOTE with auto sampling strategy	Poor
178	0.6	0.51, 0.67	0.68	0.59	0.69	Trial ≤ 3.5	Trial ≤ 1.5 -	Cost complexity pruning: alpha = 0.002	Oversampling: SMOTE with auto sampling strategy	Poor
179	0.61	0.65, 0.56	0.68	0.59	0.69	DI ≤ 1.5	Trial ≤ 5.3 DI ≤ 2.5	GridSearch CV: max depth = 5, max features = 1	Oversampling: SMOTE with auto sampling strategy	Poor
180	0.6	0.51, 0.67	0.68	0.59	0.69	Trial ≤ 3.5	Trial ≤ 1.5 -	Max depth 5	Oversampling: SMOTE with auto sampling strategy	Poor

Supplementary Table 2. Summary of the accuracy, F1-scores, ROC area under the curve, K-fold cross-validation minimum and maximum accuracy scores, the root features, 2nd split features, pruning technique, imbalanced class solution, and ranking of every decision tree produced in the *ETV6*::*RUNX1* subgroup. Decision trees were ranked as good if the F1-scores were ≥80% in each group and poor otherwise. 4 class outcome: continuing remission, died in remission, relapse/ refractory disease leading to 2nd relapse, and relapse/ refractory disease leading to death. 2 class outcome RR: relapse/ refractory yes vs no. 2 class outcome RD: remission death yes vs no. Total drug dosages: steroid, vincristine, asparaginase, methotrexate, purine, anthracycline, cytarabine, cyclophosphamide, etoposide. Acc: accuracy, CV: cross-validation, min: minimum, max: maximum, cyt: cytarabine, vin: vincristine, asp: asparaginase, meth: methotrexate, anth: anthracycline, cyc: cyclophosphamide, etop: etoposide, DI: delayed intensification, reg: regimen.