Missing from Lancashire: The Influence of Risk Assessment on Time to Resolution in Missing From Home Cases

Jessica Phoenix

31975345

MSc Quantitative Methods
for Science, Social Science
and Medicine

Abstract

In 2014/15, police forces across England, Scotland and Wales received over 300,000 calls relating to missing persons, a figure that appears to be increasing. Despite this growing figure and the workload it places on police forces, there has been a lack of research into the area of missing persons. For most forces, the level of police response that a missing person case requires is set by the risk classification that has been assigned to the case. These levels are 'Standard Risk', 'Medium Risk' and 'High Risk'. This project examines the appropriateness of these risk classifications based on how they are assigned and the effect they have on the time it takes to resolve a case. The data used comes from Lancashire Constabulary and contains all missing person reports that were made to the force in 2015. Logistic regression is used to investigate the individual risk factors that best predict a 'High Risk' classification and examine how these differ to the risk factors that the police believe indicate higher risk. The main body of the analysis focusses on modelling the time to resolution for missing from home cases as predicted by their risk level and other explanatory variables using event history analysis methods. Kaplan-Meier estimates are used to model the probability of no resolution by risk level. Cox Proportional Hazards Models are then used to determine which factors alongside risk level are significant in their prediction of time to resolution. These models are then extended to account for the effects of repeatedly missing persons with the inclusion of a frailty term. This project concludes that risk classification does have a significant effect on the time to resolution and puts forward the notion that the large amount of cases being classified as 'Medium Risk' has absorbed police time and resources and in turn taken these away from the more complex 'High Risk' cases which see a slower time to resolution than 'Medium Risk' after the first 24 hours of a missing person case being created. Recommendations and possible extensions to the analysis are provided.

Contents

List of Tables and Figures

1. Introduction

According to figures presented in the UK Missing Person's Bureau (2016) missing person data report, 321, 992 calls relating to missing persons were made to police forces in England, Wales and Scotland in 2014/15. This was calculated to be a 3% increase in calls made in comparison to 2013/14. The 2014/15 report collates data from police force command control systems and missing person case management systems. All 43 police forces in England and Wales submitted full data from each quarter on both incident and individual level data whilst Police Scotland gave annual figures for calls received relating to missing persons. The 2014/15 report reflects the first year in which the 'absent' category was clearly applied by 38 participating police forces, including Lancashire, though it is acknowledged that the consistency of this application varied by force.

The definition of a 'missing person' which is supported by Lancashire Constabulary is given by the National Police Chief's Council (NPCC), an organisation which replaced the Association of Chief Police Officers (ACPO) in 2015. They state that an individual whom can be classified as a missing person is 'anyone whose whereabouts cannot be established and where the circumstances are out of character or the context suggests the person may be subject of crime or at risk of harm to themselves of another' (ACPO, 2013: 5). An absent classification on the other hand relates to 'a person not at a place where they are expected or required to be and there is no apparent risk', as updated by ACPO in 2015. The absent category was introduced by ACPO in their 2013 guidance for managing cases of missing persons, due to the challenge faced by police forces in attempting to effectively investigate the vast volume of missing person reports that were being received. The guidance stated that the inclusion of the absent category was embedded in a new and more risk-based approach to responding to missing person reports. Those classified as absent would still be monitored, but would be dealt with in a more effective way.

In addition to being classified as either missing or absent, within the missing category are also three levels of 'risk' which are assigned as appropriate to each person classified as missing. The level of perceived risk affects the immediate police response and future monitoring of the case. Lancashire Constabulary use three levels: 'Standard Risk', 'Medium Risk' and 'High Risk'. Standard Risk is defined as a situation in which 'there is no apparent threat of danger to either the subject or the public'. Medium Risk refers to cases where 'the risk posed is likely to place the subject in danger or they are a threat to themselves or others' and High Risk relates to cases in which 'the risk posed is immediate and there are substantial grounds for believing that the subject is in danger through their own vulnerability; or may have been the victim of a serious crime; or the risk posed is immediate and there are substantial grounds for believing that the public is in danger'. The main aim of this dissertation is to analyse the relationship between these initial risk assessments given to missing person reports and the result of each case, in other words, does the assigned risk affect the time taken for a case to be deemed 'resolved'?

Also of interest is the demographics of the missing; how factors such as age, ethnic group and gender are represented in the missing population and how said factors influence the time taken to 'resolution'. The UK Missing Persons Data Report 2014/15 gave information on the age, gender and ethnicity of the missing population. It was found that despite males making up 49% of the general population, they accounted for 52% of the missing population. Of the

children that were reported missing, that is those aged between 0 – 17 years, 54% were female. Of these, 95% were aged between 12-17. The 12-17 age bracket were overall the most likely age bracket to be reported missing, accounting for 56% of total recorded incidents. In contrast, for the adult missing population, that is over 18 years of age, 62% were male. Overrepresented in this demographic were males aged between 22-39. The problem of young males going missing was further investigated by Newiss (2015) and Kingston University whom analysed data on missing adult males who were found deceased after going missing on a night out between January 2010 and August 2015. It was found that with a range of 16-62, the most common age of being reported missing was 18. Of the under 25 year olds, 35% were students. Most cases related to the winter months December, January and February and in 89% of cases the bodies were discovered in bodies of water. The report has suggested that more care be taken for males by their friends and colleagues during nights out during the festive period, particularly when around bodies of water such as rivers and canals. Regarding ethnicity in the missing population of the 2014/15 report (UK Missing Person's Bureau, 2016), people of white ethnicity made up most of missing persons at 71%. People of white ethnicity accounted for 86% of the general population and so could be considered underrepresented in the missing population. On the other hand, people of black ethnic origin made up 11.2% of the missing population despite only accounting for 3.3% of the general population and are thus overrepresented in missing person cases, though the report does not offer a reason for this. Those of Asian origin accounted for 3.3% of the missing population and 7.5% of the general population. In all but Chinese, Japanese and South Asian groups, males were most likely to be reported missing. Demographics relating to sexual orientation and gender outside of the binary were not covered by the data.

Further information on risk assessment, current police procedure and the objectives of this analysis will be given in later sections.

1.1.Previous Research

Prior to conducting the analysis, a selection of previous research was examined to understand areas of concern that have been identified, in addition to recognising previous recommendations made from earlier findings.

1.1.1. Newiss (1999) Missing Presumed…?

A dominant influence for the current project was the early work of Newiss (1999) whose research recognised the lack of previous attention to the police handling of missing person cases despite the hundreds of reports that the police received daily and high amount of resources that missing person reports consumed. Newiss examined the practices of police forces across the UK to identify best practices and highlight areas in need of revision, allowing recommendations for future police policies for handling missing reports to be made. Simple questionnaires asking details of the force's missing person policy and any local problems believed to contribute to the volume of reports were sent to 50 police forces across England, Scotland, Wales and Northern Ireland of which 46 responded. The questionnaires provided elementary information on procedure and the local area, from these nine forces were selected for further investigation with semi-structured interviews with policy makers and operational officers. Results from these forces are summarised.

Firstly, regarding the 'missing person problem', most of the reports related to people under the age of 18. The large volume of such reports occupies a great deal of police time as each report had to be investigated by police even though for most of these cases the young person will return unharmed without the requirement of police aid. Additionally, figures given for young people reported 'missing' are underestimated as a number of these reports are cancelled before they are circulated, though still using police resources. Examining reasons as to why someone goes missing, it seems that a small number of local issues contribute to large amount of missing reports. The most common reason was young people being reported missing from children's homes, referred to as 'repeat runaways' due to repeatedly being reported as missing. The duration of these cases is often short as the missing young person returns to the children's home yet they contribute a substantial amount to police workload. Police forces often saw repeat runaway reports as an administrative task as opposed to a genuine report, and all forces spoken to had either implemented a specific policy or took on a local procedure for handling repeat runaways. In addition to missing reports from children homes, of concern to the police were 'suspicious missing persons'. Such reports refer to missing persons who are at risk or have been victim of a serious crime such as homicide or abduction. Whilst these cases only make a very small minority of missing person reports, police procedures are required to differentiate between suspicious and non-suspicious cases. Of interest from the police force responses in Newiss' research is the suggestion that many suspicious missing persons are not those who have been previously classified as 'vulnerable'. This is something that will be considered in the present analysis assessing the appropriateness of initial risk assessments given to reports of missing people.

Procedures for dealing with missing person reports varied between forces, though for most the initial report was taken and assessed for priority by control room staff. This assessment determined the speed as to which police were deployed and many forces expressed concern that control room staff were not trained enough to make such assessments. Most forces then agreed that an initial search of the missing person's property was a priority, this could include searching residences of family and friends and would be more intensive for missing persons deemed 'high-category', which of course entails more police resources. The classification of missing persons for most forces was based on 'vulnerability' with only one participating force basing assessment on 'risk', though this is now recommended in national police guidance (ACPO, 2005). 'Vulnerable' has been used to refer to the young, elderly, mentally unwell, drug-dependent and long-term missing, though Newiss states this classification rarely specifies the risk that individual may be subject to. It could well be possible therefore that excessive or negligent police resources are deployed to cases due to these classifications. This not only could lead to an unnecessary drain on resources, but may lead to a high-risk case not being recognised and thus not resolved and result in negative press towards police action. Also considered by Newiss in police procedures was the time frame of handling missing reports and the circulation of these reports to external parties. Regarding practice when a missing person was found, there was limited action the police could take and only legal power when dealing with a young person. Return interviews should be conducted with all found missing people, though it was admitted that this was often not complete if the case was a repeat runaway as again this was deemed unnecessary use of police resources.

Finally, Newiss highlighted several strategic issues that had been identified throughout the research. To begin with, there was ambiguity as to who was responsible for handling the

initial missing person report. It was felt by several officers that the post of a missing person officer would be beneficial. Additionally, strategic problems arose with the recording of reports due to inconsistent systems and time consuming missing report forms. The need for more effective inter-agency working for locating missing persons was also highlighted due to the complexity of cases. For instance, Newiss referred to links between missing persons and prostitution, especially in the cases of missing from children's homes. In addition, some forces raised concerns of missing persons from ethnic minorities; particularly young girls fleeing arranged marriages or domestic violence in these marriages. These cases are complex and benefit from specialised posts.

To summarise, Newiss highlighted that most missing reports which occupy a large amount of police time and resources are those which are often resolved quickly and are deemed a routine task rather than a genuine investigation. A need for clear procedure to identify the minority cases in which missing persons are at serious risk was emphasised. Newiss made several recommendations for police forces, future research and the Home Office. It was suggested that clearer pathways need to be formed for allocating responsibility to who should deal with missing person cases and the classification criteria of missing persons and the effect this had on police response needed to be revised. The personnel, timing and terms of reference for effectively reviewing missing person reports over time needed to be agreed. It was also recommended that the systems used for recording cases be considered and accountability ensured in these records so that necessary information would be readily available should the case escalate. Also stated is that the inexperienced officers dealing with cases need supervision and support to identify suspicious missing persons as procedures for doing so were underdeveloped. Finally, Newiss recommended that future research assess the reliability of the classification criteria for missing person reports and provide clear guidance to the police on how to better identify suspicious missing persons early in the case. These final recommendations are embodied in the current project to assess the appropriateness of what is now the initial risk assessment attached to missing person reports.

1.1.2. APPGs' (2012) Joint Inquiry into Children who go Missing from Care

Due to the vast amount of cases, focus of the literature review turned to young people being reported from children's homes. Whilst Newiss' (1999) earlier research highlighted police attitudes towards reports of children missing from care as administrative tasks rather than genuine investigations, recent high-profile cases and research have demonstrated these reports as a serious cause for concern and a higher priority than perhaps previously thought. This latter viewpoint is represented in the joint inquiry conducted by two All Party Parliamentary Groups (APPG) (2012) into the management and support given to looked after children who are reported missing. The APPG for Runaway and Missing Children and Adults and the APPG for Looked After Children and Care Leavers argued that there is a scandal across England regarding looked after children going missing from care that has only started to be acknowledged due to cases such as the child sexual exploitation (CSE) in Rochdale being recently uncovered. Details on responses to the CSE in Rochdale that occurred between 2006-2013 are given in an independent reviewing officer report (www.rochdale.gov.uk/independentreview), this is following the conviction of nine men from Rochdale and Oldham for child grooming offences. The APPG inquiry collected data from those who had been missing, ministers, national agencies such as Ofsted, the voluntary sector, police forces and local authorities and examined numerous issues such as looked after

children being placed far from home, the police's response to missing care children and Ofsted's role in safeguarding these children.

The inquiry states that at the time of writing, an estimated 65,000 children were under the care system in England. Of those it was estimated that 10,000 cared for children went missing in one year, this makes children in the care system around three times more likely to be reported missing than any other children. The inquiry states that whilst missing, these children are at serious risk of physical violence, CSE and often resort to theft to survive. Therefore, a child going missing from care presents many different, complex issues. Throughout the data collection the researchers heard stories of abuse from children that had been missing. Such stories were also heard from the relevant professionals associated with children that are reported missing from care. It was found that negative and dismissive attitudes from professionals towards these children who suffer abuse whilst missing often led to such trauma going undetected. The inquiry also stated that a concerning number of vulnerable older children were being placed in poor quality care homes and that almost half of all looked after children were being allocated care many miles away from their home, family and friends, despite evidence showing that this is a causal factor in children running away from care.

Based on the findings of the inquiry, the two APPGs provided a list of key recommendations to improve the way children who go missing for care are handled to lower the number of those going missing and provide appropriate support for those who do. The six recommendations stated; a need for an independent investigation into England's children's homes that are failing to manage and protect runaway or missing children, and the introduction of a 'scorecard' for local authorities to measure the protection of missing care children. This is in addition to urgent action for the children placed outside of their local authority and a reduction in the number of children being assigned care far from home. They also argued that barriers restricting the police from information on all names and addresses of children's homes in their area need to be overcome and a new reporting system for missing incidents of children in care needs to be implemented that combines information from both the police and local authorities. Finally, it was recommended that more weighting be given to missing incidents in Ofsted's inspections of children's homes, preventing a 'good' rating being given to homes with a high number of missing children reports.

Following this inquiry, the risk assessment given to children missing from care and the resulting outcome of these cases will be examined throughout the succeeding analysis to determine if such assessments are appropriate based on if and how these children are 'found'.

1.2. Objectives

The primary objective of this dissertation is to determine whether the risk assessment given to a missing person on receipt of the initial report; that is 'Standard Risk', 'Medium Risk' or 'High Risk' is appropriate - based on the outcome of that case. The project is based solely on reports made to Lancashire Constabulary in from the period of 2015 to mid-June 2017. 'Absent' classifications are not investigated. The 'outcome' of the case refers broadly to whether that missing person was found alive or deceased, or if the case is still outstanding. Whilst definitions of 'resolution' may be contested, for this analysis a case will be deemed 'resolved' if the missing person has been located alive or deceased and deemed 'not resolved' if not found/unknown, including cases that have been 'transferred'. Cases that have been

transferred from Lancashire to a different police force have for this analysis been deemed ‘not resolved’ as the cases were no longer under the control of Lancashire and so the outcome not followed up. The probability of a case being resolved and the risk of it not will be modelled using event history analysis techniques, though more details on methods of analysis will be given in the following chapters.

Secondary objectives in the project include identifying how a case is assigned a risk level based on the risk assessment questions asked to the informant who reported the person as missing. This will indicate what it means to be ‘high risk’ under Lancashire’s current assessment procedure. It is beneficial to understand how these classifications are assigned due to the level of police response associated with each risk. This investigation was conducted prior to the main analysis so that the risk classifications were understood as best as possible before their influence on the resolution outcome was assessed. Following the primary analysis, the influence of persons repeatedly being reported missing or ‘repeat runaways’ was examined using frailty models; an extension to event history analysis. The intended outcome of this final investigation was to provide more accurate estimates of predicted resolution times with event correlation being accounted for.

Finally, this project aims to highlight the demographics of those that are most likely to be reported missing based on the proportion of reports. These are predominantly age, gender, ethnicity and care status. The towns in which the most reports are made from will also be examined. This information will act as a guide for Lancashire Constabulary to identify the areas and people that are most at risk and ensure police resources are being targeted to the areas that are most in need of monitoring and support and thus prevent increasing missing reports and reduce police demand.

Implications drawn from the results of this analysis hope to advise future policy surrounding the initial risk assessment of missing person reports to ensure police are appropriately being deployed to the higher risk cases whilst resources are conserved when dealing with a lower risk report. An accurate and justified assessment of risk when a person is reported missing would not only benefit the police in resource allocation but also ensure those most in need of support can receive it.

1.3. Current Lancashire Constabulary Procedure

Current Lancashire Constabulary guidance lays out a standard procedure for handling missing person reports that distinguishes between ‘missing’ and ‘absent’ cases and aims to ensure investigations are concentrated to the most vulnerable people and make the best use of police resources. On receipt of an initial report, the call taker would assess the risk of the missing person using a standardised list of risk assessment questions that are designed to extract detailed information, questions include information on the physical and mental health of the missing person and any personal circumstances that may be of concern. If the call taker deems a missing person to be ‘absent’ or to be of Standard Risk, the case is sent to the Demand Reduction Unit for review. If it is believed that the case is Medium Risk or High Risk, the case is sent for immediate deployment. This is a dynamic process and the risk assessment may change throughout the case time frame should new information be received.

Each classification requires a different police response. If absent, a time frame for reviewing the case is set. After 15 hours, depending on the circumstances then the case is either closed

or reclassified as missing. It is stated that no person under the age of 16 should be classified as absent for longer than this 15-hour period. For those classified as 'missing', each risk level elicits different amount of police attention. For a Standard Risk, the report is recorded onto the relevant IT systems and the informant is made aware that once all active enquiries have been addressed, the case will be left for regular review until further information is received. If the missing person is under 16, the classification is increased to Medium Risk. For a Medium Risk case, there is an active response by police to locate the missing person and support the informant. Finally, for a High Risk report there is immediate police deployment and sometimes a Senior Investigating Officer is appointed. Close contact is monitored with other relevant agencies and the press.

Regarding the return of a missing or absent person, responsibility varies on circumstance. For those classified as absent, it is the responsibility of the party who made the report to inform police of the return. There will be no return interview unless the absent case is a young person who has been deemed absent multiple times in which it could be considered by the missing from home (MFH) co coordinator. For returning missing persons, if returning to a care home it is the responsibility of care staff to return the individual and to conduct a safe and well check. If the person is returning to a foster home of family home, it is the responsibility of the police to attend for a safe and well check. The following section will outline the dataset used in the analyses that was provided by Lancashire Constabulary and the reasoning behind the variables chosen.

2. The Dataset and Initial Data Exploration

Prior to data being selected, meetings were held with Lancashire Constabulary and the Missing from Home team to discuss any current issues with the handling of missing person cases in Lancashire that could be addressed by the present research project. Highlighted in these discussions were concerns surrounding the risk classification assigned to missing persons in the initial report and the effect this had on police response and whether or not these classifications were appropriate. An inappropriate classification as noted earlier can lead to unnecessary police deployment to cases classified as medium or high risk that perhaps did not warrant this level and negligence to cases deemed standard risk which required more police attention. These discussions alongside examination of previous research have provided the basis for this analysis. The data were then provided by Lancashire Constabulary following discussion on which variables were deemed relevant and which time-period would provide enough follow-up time for a detailed analysis. Level 3 vetting and agreement of confidentiality was required before the data could be accessed. The dataset contains details on all cases of persons reported as 'missing' in Lancashire between 1st January 2015 and 31st December 2015. The original dataset contained 5952 cases, following data formatting this was reduced to a final number of 4746 observations. The dependent variable throughout the analysis was the survivor function or hazard rate, with time defined by the 'days_missing' variable, more detail on the variables will be given in the following subsection.

2.1. Data Formatting

The raw data were obtained from Lancashire Constabulary in Microsoft Excel format, from here they were made anonymous. All variables that could be used to identify a missing person were removed, this included their home address, the address of which the person was missing from, the address in which they were located and the address in which they were last

seen. Also removed were the descriptions of the missing and the return circumstances as these contained personal details surrounding the missing person, including their name and names of associates. A code was applied to a new column in Microsoft Excel to assign the name of each missing person a unique number ensuring that missing persons retained the same unique number each time they were reported missing. This allowed repeat cases to be easily identified. The edited data were sent to Lancashire Constabulary to ensure they fulfilled the appropriate level of anonymity and could be analysed, the data were approved.

Whilst still in Excel, a new variable labelled 'transferred' was created to monitor cases that had been transferred from Lancashire to a different police force. These cases were located through a search of the data for keywords 'transferred' and 'passed'; those that had been moved to another force were coded as '1' and those that had not as '0'. It is acknowledged that not all transferred cases were necessarily located due to potential misspellings when the raw data was inputted meaning they were not picked up by the search and the use of alternative keywords by officers not included in the search. A new variable was then formed labelled 'resolved' which would be the primary 'event' variable throughout the analysis, investigating how the selected explanatory variables influenced the 'time to resolution'. Cases were deemed 'resolved' if the missing person or the body of the missing person had been located and coded as '1', 'unresolved' cases were those where the outcome remained unknown and were coded as '0'. Cases that had been transferred to other forces were included in those deemed 'unresolved' as this was the point in which the report was no longer the responsibility of Lancashire and so the outcome could not be followed, observations that had been transferred were thus censored in the data. The follow up time for transferred observations will be much shorter than other unresolved cases as the observation is censored at the date of transfer. Two time variables were then formed that were necessary for the event history analysis methods used, these were variables which measured the amount of time that a person was missing for. Two existing variables were used: 'mfh_date_created' referring to the date and time in which the missing person record was created by Lancashire Constabulary and 'mfh_date_found' for the date and time that the police deemed the missing person to be located. A 'days_missing' variable was formed through subtracting the created date from the found date in a number format to give the amount of day(s) that the person was missing, this was the 'time' variable used in the analysis.

The edited data were then transferred into SPSS 23 software to facilitate a simpler way of ordering the data to identify errors and to create a new categorical age variable with desired brackets. Sorting time variables by ascending order found that several cases had been recorded has having a negative missing time-period with the found date being stated as happening before the created date. As this could not be possible these cases were assumed to be input errors and removed from the dataset, as were cases that had been recorded as missing for zero time. Cases that had been entered as duplicates were also removed. The chosen method for dealing with missing data was a 'complete case' analysis and so all cases which had any missing values in the selected variables were deleted from the dataset. The final data contained 4746 cases from the original 5952 following the above removals. An additional variable was then added through a transformation of the continuous 'mfh_age' variable to a categorical with four values representing age brackets: 1 = '0-18', 2 = '19-40', 3 = '41-64' and 4 = '65+'. This categorical age variable named 'age_cat' was formed to aid comparison between groups during the analysis. The brackets were loosely based on the

February 2016 overview of the UK population (www.ons.gov.uk) which stated the median population age in 2014 was 40 years, the lower quartile was 21 and the upper quartile 58 years. The lower middle and upper middle age brackets in the new variable contain these upper and lower quartiles. Finally, a new binary variable was created through the transformation of the existing 'mfh_risk' variable which listed each risk classification for each case. The new variable named 'BinRisk' kept all observations labelled 'High' as 'High Risk' but combined all 'Standard' and 'Medium' cases to form one 'StdMed Risk' variable, this binary variable allowed for differentiation between cases that were and were not high risk to aid the primer analysis investigating risk assessment and classification, as will be discussed in the following chapter.

The formatted data were then read into R software (version 3.4.0) and R Studio using the 'haven' (Wickham and Miller, 2017) package downloaded from CRAN. R Software was used for the remaining analyses as it has a much larger number of packages available, specifically it has the 'survival' package for conducting event history analysis which was the primary method applied to the data. A full discussion of the analysis methods used will be given in the following chapters. Nominal variables were recoded to factors and the dependent variable 'resolved' treated as numeric.

2.2. Data Exploration

Prior to investigating 'time to resolution', the data were explored to better understand the demographics of those who were missing from Lancashire in 2015 and relationships between the variables. Frequency tables and cross tabulations were produced in SPSS to observe the people and areas that encompass the majority of missing from home cases and how these relate to their assigned risk classifications.

2.2.1. Who is Reported Missing?

Frequency tables and cross tabulations in SPSS of variables relating to person age, gender, ethnicity and home address provided the demographics of the groups most likely to have been reported missing in 2015. Age and sex were initially examined and are presented in Table 1, row percentages and total percentages and shown in the brackets.

Table 1: Age and Sex of Missing Population

		Sex		
		Female	**Male**	**Total (% of Total)**
Age Bracket	0-18 (% of Bracket/% of Total)	1628 (47.9%/34.3%)	1769 (52.1%/37.3%)	3397 (71.6%)
	19-40 (% of Bracket/% of Total)	274 (35.1%/5.8%)	507 (64.9%/10.7%)	781 (16.5%)
	41-64 (% of Bracket/% of Total)	128 (31.8%/2.7%)	274 (68.2%/5.8%)	402 (8.5%)
	65+ (% of Bracket/% of Total)	58 (34.9%/1.2%)	108 (65.1%/2.3%)	166 (3.5%)
Total (% of Total)		2088 (44%)	2658 (56%)	4746 (100%)

Those aged 18 years and under largely make up the majority of missing person cases and account for 71.6%, of these 37.3% are male. In each age bracket there are a larger number of males than females that have been reported missing, with the biggest gender gap in the 19-40 age bracket. A look at the continuous age variable found the most common single age to be 15 years which accounted for 20.4% of all missing cases, closely followed by 14 with 18.1% of cases.

Sex and ethnicity were then compared. Overall, those identifying as White British made up the large majority of cases at 92.2%. This could suggest overrepresentation when compared to the most recent Ethnicity and National Identity publication from the Office of National Statistics (ONS, 2011) which found White British people accounted for 80.5% of the general UK population. Other groups that may have been overrepresented were missing people of mixed White and Black Caribbean background at 1.1% whilst making up between only 0.5 – 1% of the general population and any other Mixed Background which accounted for 0.7% of the missing population and 0.5% of the general UK population. It is acknowledged that the numbers of ethnic backgrounds in the UK may have changed between the time of the 2011 publication and the collection of the 2015 data. Males made up the majority within each ethnic group excluding Pakistani, any other Asian Background, African and Mixed White and Asian in which females were more likely to have been reported missing. The same percentage was found for both females and males in Indian, Mixed White and Black Caribbean, Mixed White and Black African and Irish groups. The least likely ethnic backgrounds to be present in the data were 'Any Other Ethnic Group' in which there were no cases and 'Chinese' which had only one case.

Frequency tables of the locations involved in MFH cases in Lancashire found that Blackpool and Preston were the most common home towns of missing persons and the most common towns in which persons were reported missing from. Blackpool accounted for 17.4% of home towns and 18.1% of places the person was missing from. 16% of home towns related to Preston and 16.3% of missing from locations were also Preston. A cross tabulation of home locations and age found that 12.6% of all MFH cases related to persons aged 18 and under from Blackpool and a further 11.3% were aged 18 and under from Preston. A further look found that of the 16.3% of missing cases from Preston, almost half (7.1%) related to children and youth under local authority care. 6% of Blackpool cases also related to children and youth under local authority care whilst the remaining 6.8% were accounted for by any other missing child under 18.

Following these frequency results, more attention was paid to the relationships between age and other variables in the data, starting with care status. The majority of missing persons were not under any care order (61.6%), though a large proportion of reports did relate to children and youth under care orders. 18.5% of the MFH cases related to persons aged 18 and under accommodated by any local authority under Section 20 Children's Act 1989 and a further 11.6% were accounted for by aged 18 and under persons under Section 31 Children's Act 1989 Care Order.

With regards to the end of the MFH case in which the person was located, 44.7% returned of their own accord. These were most likely to be those between ages 0-18 and 19-40 whilst the older age brackets 41-64 and 65+ were most likely to be found by the police. Missing persons in the 41-64 age bracket were the most likely to be found in hospital or deceased. The

youngest bracket of 0-18 were the most common to have an 'unknown/other' outcome accounting for 3.1% of cases, however it is known that almost a third of these relate to cases that were transferred to other police forces and so the outcome was not followed up by Lancashire. Those aged 0-18 held the largest percentage in all return outcomes excluding those found in custody, hospital or a hotel/other commercial premises in which the most likely group was 41-64. Regarding where missing persons stayed whilst missing, 43.2% of the sample stayed with friends and for 33.2% their location whilst missing remains unknown. Of all the MFH cases, less than 10% were deemed 'unlikely' to go missing again. This suggests an expectation of the same persons being reported missing to the police repeatedly.

2.2.2. Which Cases are Transferred?

A closer look was given to the cases that are transferred to other police forces, as whilst they are not followed by Lancashire until the closure of the case and the outcome therefore often remains unknown, they still contribute to Lancashire police work whilst they are monitored until their date of transfer. In total, 55 cases were transferred to other forces, accounting for 1.2% of all reports. The gender split of these observations found cases relating to males were more likely to be passed to another force as they accounted for 61.8% of all transferred cases. Regarding age, the majority of transferred cases related to children and youth between the ages 0-18, accounting for 49/55 cases (89.1%). There were 3 cases passed to another force relating to persons between 19-40 and 3 cases relating to persons between ages 41-64. Of the transferred cases relating to children and youth, a large majority were those under some form of local authority care, as shown in Table 2.

Table 2: Transferred Cases by Missing Status

		Transferred	Not Transferred	Total (% of Total)
Missing Status	Adult Missing From Hospital (% of Category/% of Total)	2 (1.1%/0.0%)	178 (98.9%/3.8%)	180 (3.8%)
	Missing Child/Youth (Under 18 Years) – Cared for by Any Local Authority (% of Category/% of Total)	46 (2.3%/1.0%)	1937 (97.7%/40.8%)	1983 (41.8%)
	Other Missing Adult (% of Category/% of Total)	4 (0.3%/0.1%)	1210 (99.7%/25.5%)	1214 (25.6%)
	Other Missing Child/Youth (Under 18 Years) (% of Category/% of Total)	3 (0.2%/0.1%)	1366 (99.8%/28.8%)	1369 (28.8%)
Total (% of Total)		55 (1.2%)	4691 (98.8%)	4746 (100%)

In summary it can be seen that most transferred cases are those relating to missing youth below the age of 18 who are under local authority care, making up 83.6% of all cases passed to other police forces. Based on findings from APPG's (2012) inquiry as discussed in the first chapter, it could be sensible to assume these cases may relate to children who have been placed in care far away from their original home, often due to lack of availability, a factor

found to be causal in young people running away. These cases may be therefore transferred to the police force of the cared for child's home area in which they have more connections or family and are likely to be whilst missing. This can often lead to issues between police forces when deciding who should take control of a case. Taking the focus away from age and status, it was found that the majority of cases that were passed to other forces were those classified as 'Medium Risk' with 85.5% of all transferred cases being classified as such. This figure reflects the overall proportions of risk assignments for all MFH cases in the data, as will be discussed in the following subsection.

2.2.3. Who is most at 'Risk'?

Cross tabulations were then produced in SPSS to examine how the demographics of the missing person may relate to the initial risk level assigned to them at the time of reporting. 83.1% of cases were classified as medium risk, this was the most common classification across each age bracket. This was the large majority for cases of aged 0-18 at 90.5%, only 3.8% of cases for this age group were deemed standard risk. Missing persons aged 65+ were the most likely age bracket to be classified as high risk relative to their number of reports, with 45.8% classified as such. This is very high compared to all cases in which 10.2% were deemed high risk. Only 5.7% of 0-18 year old missing persons were assigned high risk, this was 16.2% for the 19-40 and 22% for the 41-64 bracket. A comparison of risk levels according to missing status is presented in Table 3.

Table 3: Risk Level and Missing Status of Missing Population

		Risk Level			
		High	**Medium**	**Standard**	**Total (% of Total)**
Missing Status	Adult Missing From Hospital (% of Category/% of Total)	21 (11.7%/0.4%)	152 (84.4%/3.2%)	7 (3.9%/0.1%)	180 (3.8%)
	Missing Child/Youth (Under 18 Years) – Cared for by Any Local Authority (% of Category/% of Total)	99 (5%/2.1%)	1824 (92%/38.4%)	60 (3%/1.3%)	1983 (41.8%)
	Other Missing Adult (% of Category/% of Total)	277 (22.8%/5.8%)	745 (61.4%/15.7%)	192 (15.8%/4%)	1214 (25.6%)
	Other Missing Child/Youth (Under 18 Years) (% of Category/% of Total)	86 (6.3%/1.8%)	1222 (89.3%/25.7%)	61 (4.5%/1.3%)	1369 (28.8%)
Total (% of Total)		483 (10.2%)	3943 (83.1%)	320 (6.7%)	4746 (100%)

The table shows that almost 40% of missing person cases related to a medium risk missing child under the care of any local authority. This is followed by 25.7% of cases relating to any other child of medium risk, the third most common being a medium risk missing adult not missing from hospital accounting for 15.7%. To summarise the data exploration, it is

apparent that the bulk of missing person reports in Lancashire relate to children and youth aged 0-18, in particular children and youth under local authority care. Many of these reports are concentrated in the areas of Blackpool and Preston. These cases tend to be of 'medium risk' and so require the relevant police response to these risk level cases. It is also seen that whilst cases relating to person aged 65 and over are the minority, they are more likely to be deemed 'high risk' and thus may require more police attention. The following section investigates what is meant by 'high risk'.

3. Risk Prediction

As the main aim of the project was to investigate the influence of the initial risk assessment on the time to resolution in MFH cases, it was important to firstly examine what is meant by each risk classification. In particular, it was of interest to understand which factors lead to a 'High Risk' classification as these cases require the most intense police response with regards to time and resources. An analysis into the factors which are most influential in determining a high-risk classification was therefore conducted prior to investigation time to resolution.

As discussed earlier, part of Lancashire Constabulary procedure on receiving a report of a missing person is to ask the informant a set of risk assessment questions to determine any risks that the person may be subject to whilst missing. The answers to these risk assessment questions for each of the cases are included in the MFH dataset as 19 individual risk factor variables. Each of the variables has a binary outcome of 'Y' for yes this missing person is at risk and 'N' for no this person is not at risk. These risk factor variables were used in this analysis to determine the important predictors of a high risk classification. The list and description of these variables are given in Table 4.

Table 4: Risk Factor Variables

Variable Name	Description
Mfh_risk_factor1	Is the person vulnerable due to age or infirmity or any other similar factor? ***
Mfh_risk_factor2	Behaviour that is out of character is often a strong indicator of risk; are the circumstances of going missing different from normal behaviour patterns?
Mfh_risk_factor3	Is the person suspected to be subject of a significant crime in progress e.g. abduction? ***
Mfh_risk_factor4	Is there any indication that the person is likely to commit suicide? ***
Mfh_risk_factor5	Is there a reason for the person to go missing?
Mfh_risk_factor6	Are there any indications that preparations have been made for absence?
Mfh_risk_factor7	What was the person intending to do when last seen? (e.g. going to the shops or catching a bus) and did they fail to complete their intentions?
Mfh_risk_factor8	Family / relationship problems or recent history of family conflict / abuse?
Mfh_risk_factor9	Are they the victim or perpetrator of domestic violence?
Mfh_risk_factor10	Does the missing person have any physical illness or mental health issue?
Mfh_risk_factor11	Are they on the Child Protection Register? ***
Mfh_risk_factor12	Previously disappeared and suffered or was exposed to harm?

Mfh_risk_factor13	Belief that the person may not have the physical ability to interact safely with others or an unknown environment? ***
Mfh_risk_factor14	Do they need essential medication that is not likely to be available to them? ***
Mfh_risk_factor15	Ongoing bullying or harassment e.g. racial, sexual, homophobic etc. or local community concerns or cultural issues?
Mfh_risk_factor16	Were they involved in a violent and / or racist incident immediately prior to disappearance? ***
Mfh_risk_factor17	School / college / university / employment or financial problems?
Mfh_risk_factor18	Drug or alcohol dependency?
Mfh_risk_factor19	Other unlisted factors which the officer or supervisor considers should influence risk assessment?

The star symbols attached to several of the risk factor descriptions have been placed by police as an indicator of higher risk and should thus lead to a higher risk score for that case.

3.1. Binary Logistic Regression

To determine the most important predictors of a high risk missing person, a binary logistic regression was fit treating the binary risk variable 'BinRisk' as the response to be predicted by the 19 risk factor variables which were treated as the explanatory. A binary logistic regression method was chosen due to the response variable containing only two outcomes 'High Risk' or 'StdMed Risk'. A binary outcome such as this follows a Bernoulli distribution. This distribution treats each observation as an independent Bernoulli trial and counts the number of what are termed 'successes' which in this analysis refers to a 'High Risk' classification. The 'success' is coded as '1' and 'failures' or non-high risk classifications as '0'. The aim of the logistic regression, as with standard linear regression, is to find the best fitting model that describes the relationship between the response variable and the predictor or explanatory variables (Hosmer and Lemeshow, 2000). The notation for the model can be written as:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots \beta_k x_{ki}$$

$$\text{with } y_i \sim Bernoulli(p_i)$$

The first line of the notation refers to the model and the second the distribution that the model takes. In the expression, p_i is the 'success' probability for observation *i*, in other words it is the probability of being given a high risk classification for each individual case. On the first line, β_0 relates to the intercept, then $\beta_1 x_{1i}$ relates to the first explanatory covariate for case *i*, followed by all the remaining explanatory covariates. The second line shows that the binary response y_i follows the Bernoulli distribution.

3.1.1. Model Fitting

All regression models were fit using R software and R Studio. The first model fit was a null model, estimating the ability of only the intercept in predicting the response variable 'BinRisk'. This gave a null deviance of 3122.4 on 4745 degrees of freedom which was used as a basis of comparison for the remaining models. The chosen method of selecting

significant terms to be included in the main effects model was backwards selection, this meant that the first model fitted after the null was the full model including all 19 risk factor variables as predicting the response and non-significant terms were deleted one by one. Within the model, the binomial distribution family and the logit link were specified. Once the full model was fitted, an analysis of variance (anova) was performed to identify any non-significant terms. This was done using the 'Anova' function under the 'car' library package (Fox and Weisburg, 2011) downloaded from CRAN which performs a Type Two test. The least significant term was selected based on the largest p-value, this meant that from the first anova, 'mfh_risk_factor6' relating to indications the missing person had prepared for absence was removed due to having the largest p-value of over 0.9 which was above 0.05 and thus not significant at the chosen 5% level. The following model was fitted without this risk factor and another anova ran to identify further non-significant terms. This process was repeated until a main effects model of only significant terms was left.

The main effects model found the following variables to be significant in their prediction of the binary risk variable: 'mfh_risk_factor2', 'mfh_risk_factor3', 'mfh_risk_factor4', 'mfh_risk_factor10', 'mfh_risk_factor11', 'mfh_risk_factor13', 'mfh_risk_factor17', 'mfh_risk_factor18' and 'mfh_risk_factor19'. This can be better understood as the most influential factors when assigning a high risk classification from the informant's report are behaviour that is out of character for the missing person, suspicion that the missing person is subject to a serious crime such as abduction, indications that the person is likely to commit suicide, if the person has physical health or mental illness issues, if the person is on the Child Protection Register, a belief that the person cannot interact safely with others or an unknown environment, any financial, education or employment related problems that the missing person may have, if the person has a drug or alcohol dependency and any other unlisted factors that the responding officer feels important to the risk assessment. All risk factors were found to be highly significant at the 0.1% level excluding risk factor 11 relating to the missing person being on the Child Protection Register which was still highly significant at the 1% level and risk factor 17 relating to financial, education and employment problems which was still significant at the chosen 5% level.

The final model had a residual deviance of 2568.1 on 4736 degrees of freedom, therefore a large change in deviance of 553.9 on 9 degrees of freedom compared to the null model. A log likelihood ratio test was performed to compare the two models and check for goodness-of-fit using the 'lrtest' function from the 'lmtest' library package (Hothorn et al, 2017). The test found the difference between the null and fitted model to be highly significant. The null model had a lower log likelihood of -1561.2 compared to the fitted model with -1277.6 suggesting that the final model if significant covariates was the better fit. The Hosmer-Lemeshow test was used to test overall goodness-of-fit for the model. The test was performed using the 'hoslem.test' function under the 'ResourceSelection' package (Subhash et al, 2017). This gave a p-value of 0.137 which could not reject the null hypothesis at the 5% level that the model was well fitting and so gave no evidence of a poor fit.

As this was a logistic regression, the exponents of the coefficient estimates were calculated to interpret the size of the effect that each risk factor had on the high-risk classification. Both the estimate and the exponential of the estimate are presented alongside the confidence intervals in Table 5, the first column gives a summary of each significant risk factor. All estimates have been rounded to two decimal places.

Table 5: Regression Estimates for Fitted Model of Risk Factors

			95% Confidence Interval	
Risk Factor: Reference Category (Yes)	**Estimate**	**Exponent of Estimate**	**Lower Bound**	**Upper Bound**
2: Behaviour that is out of character	1.13	3.10	2.50	3.84
3: Suspected to be subject to a significant crime	0.83	2.29	1.61	3.21
4: Indication of suicide	1.98	7.28	5.63	9.42
10: Physical Illness or Mental Health Issue	0.42	1.52	1.21	1.90
11: On the Child Protection Register	0.52	1.67	1.17	2.35
13: Belief that the person may not have the ability to interact safely with others or an unknown environment	1.03	2.80	2.05	3.80
17: Financial, education or employment problems	-0.38	0.69	0.48	0.97
18: Drug or alcohol dependency	-0.67	0.51	0.36	0.73
19: Any other unlisted factors thought to be important to risk assessment	0.72	2.05	1.60	2.63

As the confidence intervals are also presented as the exponential values, a value of 1 between the upper and lower bound would suggest no statistical significance between the groups, which in this case the groups are those who do and do not have the risk factor present. The reference category for the risk factors is 'Y' or 'Yes', meaning that the values of the odds are relating to if the person is at risk of the factor in comparison to if they were not, with all other factors held constant. The second column of exponential estimates is the main focus for interpretation. For example, from the table it can be seen that a missing person whose behaviour is deemed to be out of character is 3 times more likely to be classified as high risk than if it was not, based on the estimate of 3.10. It can also be interpreted by increase, for instance if a missing person is suspected to be the victim of a significant crime, the odds of them being classified as high risk increases by 129%, based on the estimate of 2.29. It can be seen that the largest increase of risk relates to a person who has indicated they are likely to commit suicide, this increases the odds of a high-risk classification by 628%. This was the most statistically significant risk factor in its prediction of high risk classification alongside the risk of behaviour that is out of character; both predictors were found to have the lowest p-value in the model. Conversely, it can be seen from the table that two of the factors reduce the odds of a missing person being classified high risk. If the person is believed to have financial, education or employment problems, their odds of being assigned high risk reduces by 31%. If a missing person is found to be drug or alcohol dependent, the odds of a high-risk classification reduce by 49%.

To present some of these estimates visually, a plot was produced from the final model, highlighting the two most significant risk factors: behaviour that is out of character and indication that the missing person is likely to commit suicide. The plot was produced using package ‘ggplot2’ (Wickham and Chang, 2016).

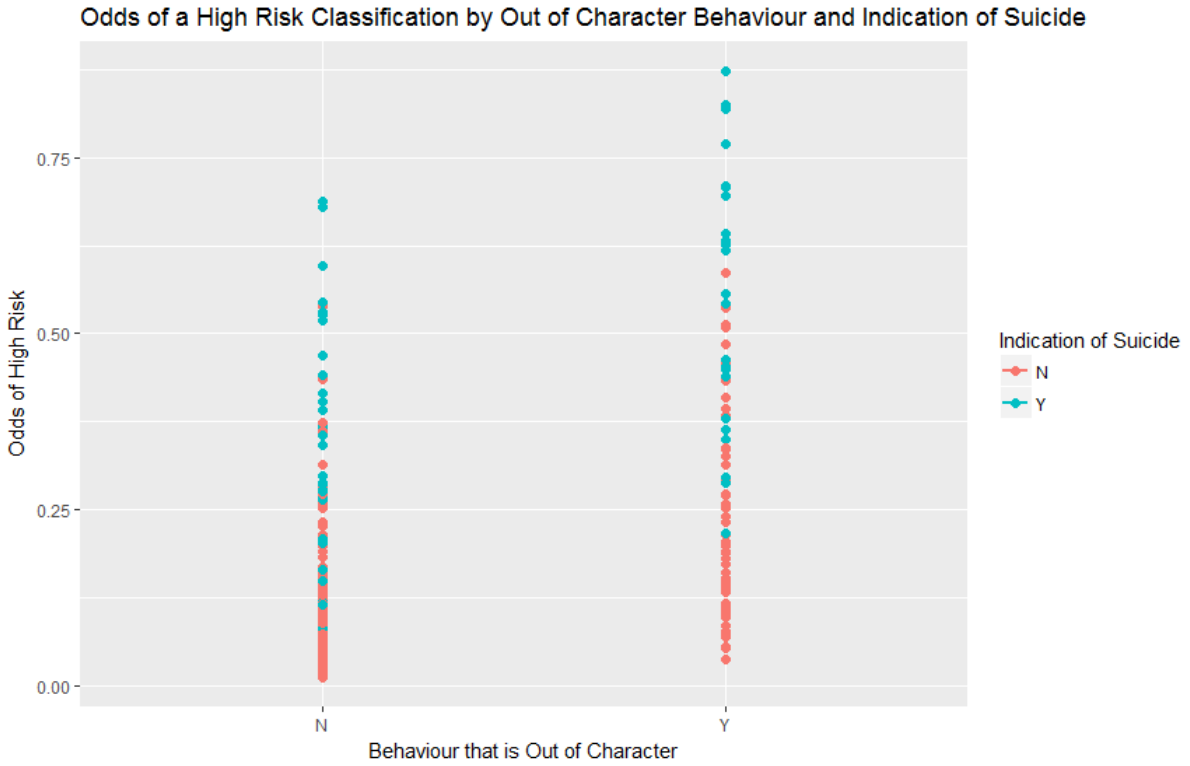

Figure 1: Odds of High Risk Classification by Out of Character Behaviour and Suicide Indication

Within the figure, ‘Y’ can be taken to mean yes and ‘N’ can be taken to mean no. It can be seen that a missing person with an indication of committing suicide has the highest odds of being deemed high risk, this appears to increase further if that person is also believed to be acting out of their usual behaviour. Additional plots were examined for the indication of suicide alongside the factors relating to alcohol and drug dependency and physical/mental health issues.

3.1.2. Interaction Effects

Following the plotting of the risk factors within the fitted model, it became of interest to investigate any possible interaction effects between risk factors. Initially, an interaction term between behaviour that is out of character (mfh_risk_factor2) and indication of suicide (mfh_risk_factor4) was added to the fitted model based on the results from the plot. An anova of this model using ‘Anova’ from the ‘car’ library which performs a type 2 test showed that this interaction was not in fact significant and so was removed. An interaction term was then added between indication of suicide and financial, educational or employment problems, this was found to be significant and so remained in the model. Following this, an interaction term was added between physical or mental illness and drug or alcohol dependency, this was also found to be significant and so remained in the model. Finally of interest, an interaction between being on the Child Protection Register and a suspicion that the missing person is subject to a significant crime such as abduction was added. Anova found this interaction to also be significant as were all the remaining covariates. A log likelihood ratio test between

this model and main effects model suggested the difference between the two was highly significant, the likelihood for the model with interactions was higher at -1277.6 than that without at -1284.2 and so was deemed a better fit. The Hosmer Lemeshow test was again used to test goodness-of-fit. The test gave a p-value of 0.347 and so gave no evidence to suggest a poor fitting model.

Table 6 shows the estimates from the main effects model with interactions. The odds are displayed as in Table 5, alongside the exponents and the exponeniated confidence intervals.

Table 6: Estimates for Main Effects Model of Risk Factors with Interaction Terms

			95% Confidence Interval	
Risk Factor: Reference Category (Yes)	**Estimate**	**Exponent of Estimate**	**Lower Bound**	**Upper Bound**
2: Behaviour that is out of character	1.111	3.03	2.44	3.76
3: Suspected to be subject to a significant crime	0.62	1.86	1.20	2.80
4: Indication of suicide	1.88	6.53	4.97	8.61
10: Physical Illness or Mental Health Issue	0.49	1.63	1.29	2.11
11: On the Child Protection Register	0.29	1.33	0.85	2.02
13: Belief that the person may not have the ability to interact safely with others or an unknown environment	1.00	2.72	1.99	3.70
17: Financial, education or employment problems	-0.72	0.49	0.28	0.79
18: Drug or alcohol dependency	-0.25	0.78	0.46	1.26
19: Any other unlisted factors thought to be important to risk assessment	0.74	2.09	1.63	2.67
4. Indication of Suicide: 17. Fincancial/Other Problems	0.83	2.30	1.09	4.99
10. Physical/Mental Health Problems: 18. Drug/Alcohol Dependency	-0.76	0.47	0.23	0.94
3. Suspected to be Subject of Crime: 11. Child Protection Register	0.81	2.25	1.04	4.90

The exponents for the individual risk factors can be interpreted in the same way as for the main effects model, for instance risk factor 3 now indicates that if a missing person is suspected to be subject to a significant crime they are 86% more likely to be classified as high risk than if they did not have this risk factor present.

For the interactions, looking at the non-exponentiated estimates, the positive estimates indicate that as the odds for the first risk factor in the interaction increase, the odds for the second also increase. For instance, the estimate suggests that as the odds of a missing person who has indicated that they are likely to commit suicide being classified as high-risk increase, the odds of financial, education or employment problems being classified as high-risk also increase. In other words, a missing person with financial, education or employment problems has higher odds of being classified as high risk if there are also indications that they are going to commit suicide. A negative estimate indicates that as the first risk increases, the second decreases. To interpret the size of the prediction for the interactions, the estimates for the individual risk factors in the interaction and the estimate for the interaction were added together and the result was then exponentiated. For instance, the effect of having both risk factor 4 and risk factor 17 present in comparison to having neither present was calculated through:

$$1.88 - 0.72 + 0.83 = 1.99$$
$$\exp(1.99) = 7.32$$

This can be understood as to have both an indication of suicide and financial, education or employment problems multiplies the odds of being given a high-risk classification by 7.32, or the missing person is 632% more likely to be high risk than if they had neither of these risk factors. This was repeated for the remaining interactions. It was found that to have both physical or mental health problems and to have a drug or alcohol dependency reduced the odds of a high-risk classification by 41% than if the missing person had neither of these risk factors, based on a calculated estimate of 0.59. To be both suspected to be subject to a significant crime such as abduction and on the Child Protection Register multiplied the odds of high-risk assignment by 5.58, in other words they are 458% more likely to be high-risk than if they had neither of these risk factors.

This model aims to demonstrate how the interaction of multiple risk factors alongside independent risk factors influence the initial assessment given to a MFH case, though there are numerous further interaction combinations that could be examined.

3.2. Summary of Risk Classification

Based on the fitted model of significant risk factors without interactions, a ‘High Risk’ classification is most likely assigned by seven of the 19 risk factors. Treated as individual predictors and with summarised descriptions, a missing person is more likely to be high risk if they are acting in a way outside of their usual behaviour, if they are suspected to be subject to a significant crime, if there is an indication that they are likely to commit suicide, if they have a physical or mental health problem, if they are on the Child Protection Register, if they are believed to be unable to interact safely with others or an unknown environment or if they are at risk of any other unlisted factor that is deemed important by the responding officer. The most influential predictors of a high risk missing person is if there is an indication of suicide followed by behaviour that is deemed out of their usual character. Significant influences on a person not being classified as high risk is if that person is believed to have financial, education or employment problems or if that person is drug or alcohol dependent. Whilst these significant factors decrease the likelihood of a high risk classification, they may be important in distinguishing between a high and medium classification, or a high and standard. It is interesting to note that these risks found to be significant to a high risk classification

differ to those that have been indicated by police as important high risk factors, identified by the star symbols in the risk factor descriptions. It may be that whilst these are indicators of high risk, in practice they are not treated as such.

The results from the fitted models allowed better understanding of which factors define a 'High Risk' missing person. Whilst those not found to be significant may not be deemed important in defining high risk cases, they may be significant in the assignment of medium or standard cases, though this is not covered here. The understanding gained from this analysis will be used in the following analyses which focus on the relationship between the three current risk levels of 'Standard', 'Medium' and 'High' and the time to case resolution. The investigation into interaction effects showed the importance of different risk factor combinations in assigning risk classification, for instance a person's likelihood of being high risk is increased even more so if they are on the Child Protection Register. An additional analysis would look further into the various possible interaction effects, though the project now turns focus to predicting time to resolution.

4. Event History Analysis

The primary aim of this project is to investigate the 'time to resolution' for MFH cases in Lancashire, particularly as influenced by the initial risk assessment assigned at time of reporting. The chosen method to investigate this was event history analysis, also sometimes referred to as survival, duration or reliability analysis depending on the field of study (Lewis-Beck et al, 2004). The following section will explain the principles of event history analysis and how these fit the MFH data.

Event history analysis is a method for longitudinal data that allows measurement of both if and when an 'event' occurs. Lewis-Beck et al (2004) refer to an event as the occasion when an observation changes from one state to another, in the case of this analysis the event would refer to a missing person being located or the case being deemed 'resolved'. A benefit of event history analysis is that it accounts for 'censoring' in the data. 'Censored' data refers to observations in which the event of interest has not occurred during the study time frame. In the MFH data, censored observations are classed as cases that were not resolved by the end of 2015; it is known that the resolution time is greater than the censoring time but the actual time to resolution is not known. This is known as 'right-censoring'. Censoring differs to observations which simply do not appear in the data which is often referred to as truncation. Flynn (2012) states that whilst the subject may not have experienced the event of interest in the study period, their 'participation' is still very important. Flynn (2012) also emphasises that whilst essential, censored observations must be assumed as 'non-informative'. Mills (2011) provides a comparison of analyses which have and have not accounted for censoring, and shows how ignoring these observations can lead to inaccurate and biased results and distorted distributions of the data. Mills (2011) expands on the benefits of event history analysis and states that it furthers the work of ordinary regression not only in its inclusion of censoring, but also in that it does not simply focus on an 'outcome' such as death or MFH case resolution, it additionally includes the time of the event. This allows comparisons to be observed between the timings of different groups, such as the time to resolution of each risk level given to missing person reports. The effects of covariates on these timings can also be investigated. In addition, event history analysis has the ability to account for time-varying covariates in which the risk of experiencing the event changes over time though this will not

be considered in the present analysis as it is not possible to measure the missing person whilst they are missing.

Mills (2011) states that the core aspects of event history analysis are the survivor function and the hazard rate. The 'survivor function' in the MFH context represents the probability that the 'resolution time' denoted *T* is equal to or greater than a specific time point, denoted *t*. Mills (2011: 9) expresses the survivor function as:

$$\hat{S}(t) = 1 - F(t) = \Pr(T \geq t)$$

$\hat{S}(t)$ represents the proportion of missing persons whose cases remained unresolved at the end of 2015. $F(t)$ is the cumulative density function and $\Pr(T \geq t)$ shows the probability of resolution time being equal to or greater than a certain time point. This function should decrease over time as more MFH cases are expected to be resolved. On the other hand, the measurement of an event occurring, for instance the probability of a case being resolved as oppose to remaining unresolved is known as the hazard rate. The main difference is that the survivor function focusses on the chances of not experiencing the event whilst the hazard rate focusses on the risk of experiencing the 'event', in this case a case being resolved. The hazard is expressed by Mills (2011: 9) as:

$$h(t) = \frac{f(t)}{\hat{S}(t)}$$

The hazard rate is summarised as the rate at which cases are resolved at time *t* given that they were unresolved at time *t.* This is therefore what Mills terms a 'conditional failure rate' (2011: 9) where each resolution is deemed a 'failure', though in the MFH data this terminology can be deemed inappropriate. The conditional rate can be written as (2011: 9):

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

The rate $h(t)$ is the instantaneous risk that the case is resolved in the time interval denoted $[t, t + \Delta t]$ given no resolution at or beyond time *t.*

Within event history analysis there are several approaches depending on the data and the desired output. These approaches refer to non-parametric, semi-parametric and parametric models, all of which have their own strengths and limitations. Non-parametric methods make no assumption about the shape of the hazard function and how this shape could be affected by different covariates, for instance in the MFH data how the shape of resolution could be affected by age, sex or ethnicity. A commonly used non-parametric method is the Kaplan-Meier estimator (Kaplan & Meier, 1958) which is a method used to observe the number of cases remaining in the cohort at a specific time point, as well as the cumulative number of cases that have been resolved at that time point (Flynn, 2012). The benefits of non-parametric methods include them providing a good descriptive basis of the data, and being relatively simple to conduct. The Kaplan-Meier estimates can be produced as plots of the 'survival curve' which is easy to interpret; curves of more than one group in the data can be also plotted and the difference between them visually compared (Flynn, 2012). The difference between groups can then be statistically compared through methods such as the log-rank test to test whether the difference is significant based on the p-value. However, a drawback of this method is that whilst the difference between groups can be tested, the actual effect size

remains unknown. Furthermore, non-parametric methods are limited in the number of groups in the data that they can compare and they cannot include the effects of multiple covariates (Mills, 2011). Despite this, these models provide a good starting point of descriptive analysis. Whilst they make no assumption of the distribution, the produced estimates and graphs can be used to choose a distribution (Le, 1997).

Such limitations can be solved using semi-parametric models, or 'multivariate survival models' (Flynn, 2012: 2793). Semi-parametric models still make no assumption about the shape of the hazard, but they do make assumptions as to how the shape of the hazard is affected by covariates. The most prominent of these methods is the Cox Proportional Hazards Model (Cox, 1972). As stated by Flynn (2012), a central aspect of these models is the assumption that the hazards of multiple covariates are proportional and this is the part of the model that is parameterised. In the MFH context, this would be the assumption that the impact of going from a standard to a medium risk assessment is the same as the impact of going from a medium to a high-risk assessment. The result of the Cox model is interpreted as the hazard ratio, and If the resulting ratio is greater than 1, the variable is associated with an increased risk of experiencing that event. In this project, the case being resolved. A hazard ratio lower than 1 thus indicates a decreased risk associated with that variable. A benefit of these semi-parametric models is that they are more flexible than non-parametric and as discussed in Flynn (2012), the interpretable graphs produced by Kaplan-Meier estimates can still be replicated with adjustments for multiple covariates. Alongside the proportional hazards, there are several other assumptions made by the Cox model. For instance, it is assumed that the hazard function is constant. Both this and the proportional hazards assumption can be tested through 'log – log' plots. In addition, it is assumed that combined variables are multiplicative, so if females were twice as likely to be resolved than males, and those aged '41-64' were twice as likely than any other age bracket to be resolved, then a female between age 41-64 would be deemed to have a quadrupled chance of being resolved. Flynn (2012) states that this assumption can be tested through the modelling of interaction terms.

The common survival models such as those introduced above can be extended in various ways to suit the aims of the analysis, for instance through the addition of time-varying covariates, parametric models and frailty models. Within the MFH are numerous observations that relate to the same subject; that is persons that have repeatedly been reported missing from Lancashire in 2015, identified through the unique code applied to each repeated name in the data formatting process. The final part of the analysis will address these repeats using frailty models: survival methods for 'recurrent events'. Further details on each method of analysis and their results will be given in the succeeding sections, firstly the Kaplan-Meier estimator will be discussed and results from its fit to the data will be presented.

4.1. Kaplan-Meier Estimator

The Kaplan-Meier method estimates the survivor function, denoted $\hat{S}(t)$, at time *t.* The survivor function in the MFH context is the probability of a case not being resolved by time *t* or the probability that the time to resolution is greater than *t* (Mills, 2011).

The estimator can be derived to and expressed as (Le, 1997: 55):

$$\hat{S}(t) = \prod_{t_i \leq t} (1 - \frac{d_i}{n_i})$$

In the expression, n_i represents the number of cases at 'risk' of being resolved at that time and d_i is the number of cases that are actually resolved at that time. The formula therefore estimates the probability of a case not being resolved by time t. Mills (2011) summarises the Kaplan-Meier formula of 'survival' probability at failure time denoted t_i, which in this case is the probability of no resolution, as the probability of a case not being resolved past the previous 'failure' time $t_{(i-1)}$, multiplied by the conditional probability of going unresolved past time $t_{(i)}$ given that it was unresolved until at least time $t_{(i)}$.

As stated earlier, Kaplan-Meier estimates are often presented as a survival curve. The main relationship of interest in the MFH is that between the assigned risk level and probability of no resolution. The analysis thus started by plotting the overall survival curve for a MFH case with average covariates which was then compared to a survival curve as predicted by individual risk levels. Details of the analysis and the results are presented in the following subsection.

4.1.1. Fitting the Survival Curve

The overall survival curve was produced firstly to estimate the probability of no resolution for a MFH case with average covariates. Plots for the Kaplan-Meier curves were produced using the 'survfit' function in the 'survival' package (Thernau and Lumley, 2017) and 'autoplot' in the 'ggfortify' library package downloaded from CRAN (Horikoshi, 2017). The 'ggfortify' package produces plots for statistical methods such as survival analysis using 'ggplot2' (Wickham and Chang, 2016) and 'colorspace' (Ihaka et al, 2016); a package for mapping colour spaces. Due to most cases being resolved within the first 24 hours, the 'days_missing' variable for measuring time was plotted with a log transformation to focus on the main area of the curve, the transformation was produced onto the plot using the 'scales' package. Time points that were felt to be notable in the case timeline were then manually added as tick marks to the x axis of the plot to provide a clearer interpretation of the transformed time axis. This initial survival curve is displayed in Figure 2.

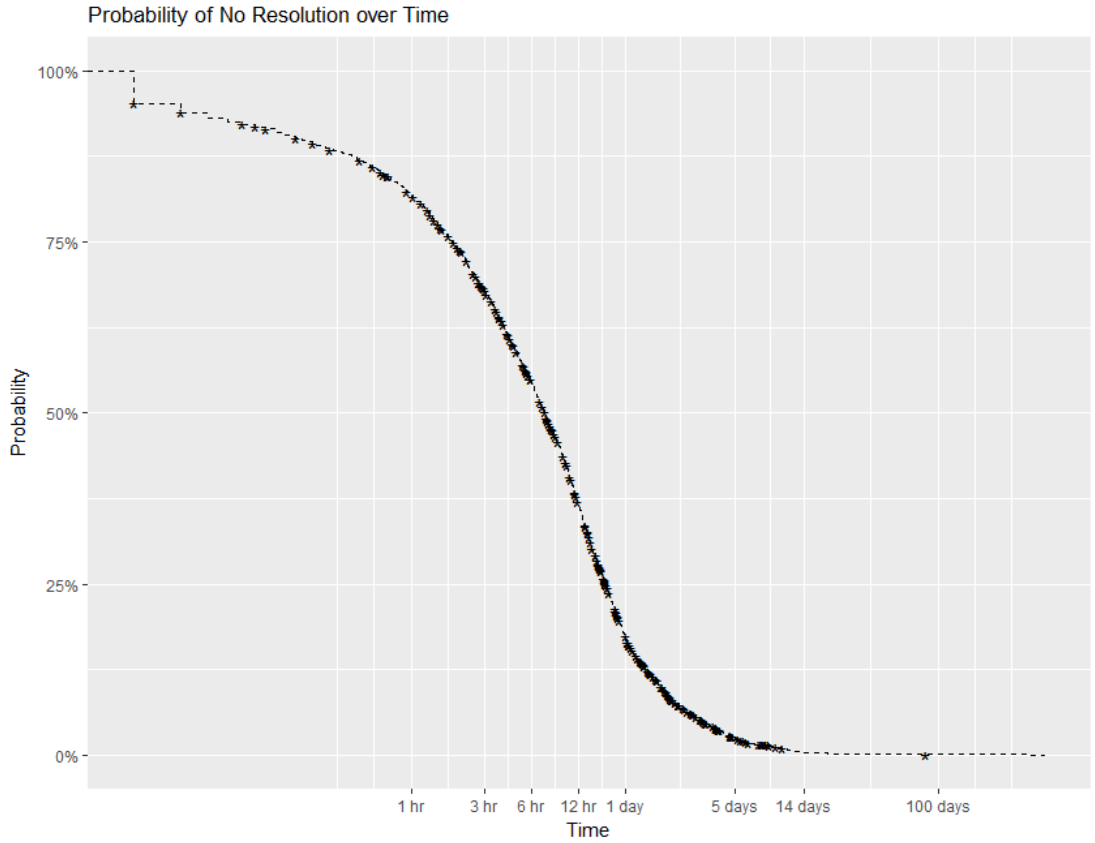

Figure 2: Kaplan-Meier Curve of No Resolution for MFH Case of Average Covariates

The y axis in the plot represents the probability of a case not reaching resolution over time. The star symbols along the curve represent censored observations. The curve suggests that half of all missing person reports are deemed resolved within 6-9 hours of being created. 75% of cases appear to each resolution between 12 hours. The majority of the remaining cases reach resolution in the first 5 days with a very small minority of cases going unresolved after 14 days. The estimate was then plotted again to show the cumulative hazard, showing an increasing curve as the hazard rate representing the risk of a case being resolved increased as oppose to the decreasing cure shown above. The estimate was then calculated for the probability of no resolution as predicted by risk level and plotted as three separate curves. Time was again log transformed and time points manually added to the x axis. The curves and their 95% confidence intervals are displayed in Figure 3.

The shaded areas around the curve represent the confidence intervals, the thinner confidence interval surrounding the medium risk curve shows that most cases are in this risk level. The curves suggest that within the first 24 hours of a report being created, high risk missing person cases are likely to reach resolution earlier than medium and standard risk, with 75% of high risk cases being resolved in the first 12 hours. After the 24-hour period the medium risk cases show a steeper decreasing curve suggesting these are most likely to reach resolution within the first one to five days of the report being created. This could mean that whilst most high-risk cases are resolved quickly perhaps due to a more intensive police response, those that are more complicated in nature and so are harder to resolve are overtaken by medium risk missing persons, which account for the large majority of cases. It could be assumed that

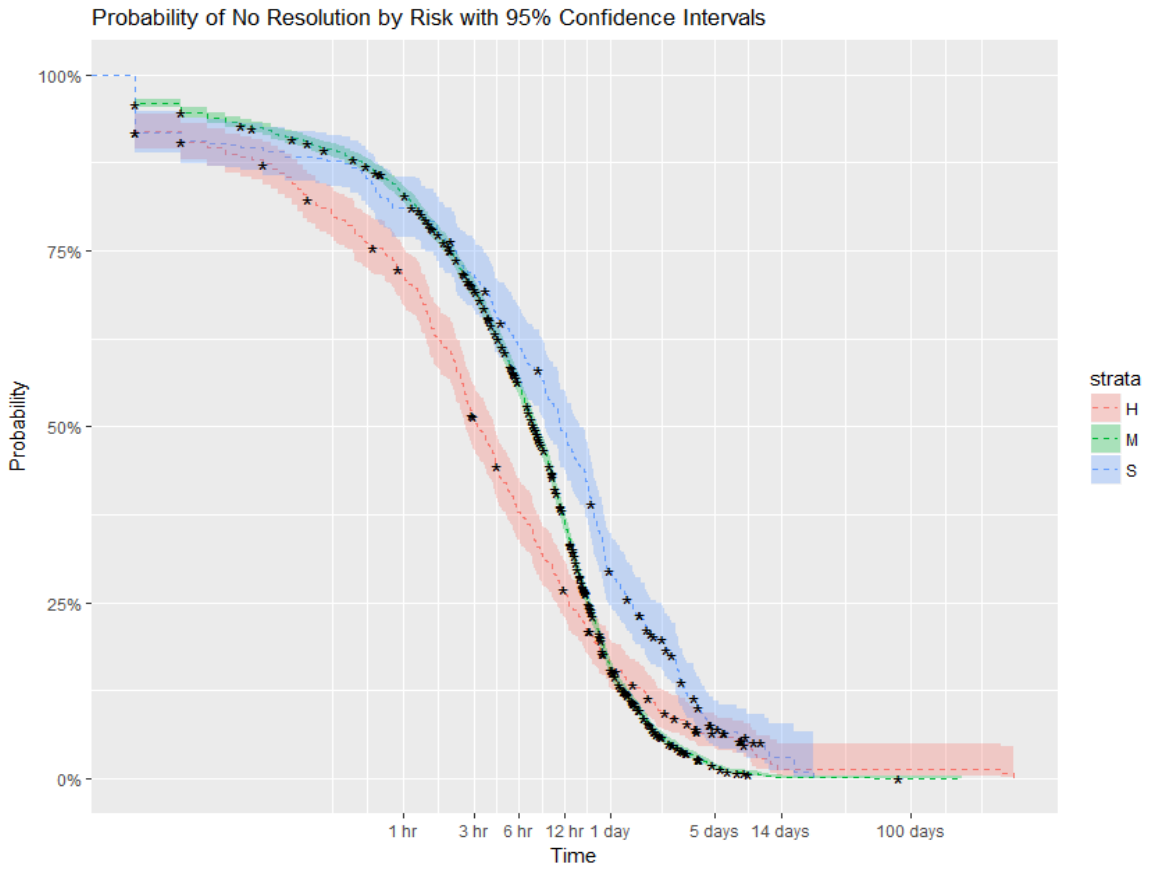

Figure 3: Kaplan-Meier Curves of No Resolution by Risk Classification

the large number of medium risk cases consume the police resources that would be needed to solve the more complicated high risks and so after the one-day period has past, the remaining high-risk cases take a greater amount of time to resolve. Standard MFH cases appear to have the slowest resolution time between the first hour and 14 days, which could be expected due to the reduced police response, though the last cases resolved in this category are done so much sooner than the two higher risk levels as shown by the curve ceasing just past the 14 day and much before the 100-day mark. To look closer at the time intervals in which notable numbers of cases are resolved by, the quantiles for each risk category were calculated. The hours and days for which 50%, 75%, 90% and 95% of cases are predicted to be resolved by as separated by risk level are given in Table 7.

Table 7: Time Taken for Cases in Each Risk Category to be Resolved

	50% Cases Resolved	**75% Cases Resolved**	**90% Cases Resolved**	**95% Cases Resolved**
High Risk	3.23 hours (2.63, 4.10 hours)	12.82 hours (10.62, 17.62 hours)	2.05 days (1.62, 3.76 days)	7.8 days (3.76, 12.93 days)
Medium Risk	7.52 hours (7.15, 8.07 hours)	17.30 hours (16.55, 18.20 hours)	1.52 days (1.41, 1.65 days)	2.54 days (2.29, 2.87 days)
Standard Risk	11.25 hours (8.90, 15.25 hours)	1.42 days (23.03 hours, 2.04 days)	3.94 days (3.08, 6.99 days)	10.74 days (4.87 days, NA)

The lower and upper bounds for the time to resolution are given in brackets below the estimate. It can be taken from the table and from the plotted Kaplan-Meier curves that risk classification does influence the time to resolution, with high risk cases reaching quicker resolution in the first 24 hours but subsequently being overtaken by the large amount of medium risk cases. The next stage was to determine whether this difference between the risk curves was statistically significant, done using a log-rank test. There are several methods for testing this difference, though log-rank as noted by Mills (2011) is the most commonly used. Mills (2011) states that these tests are calculated based on a contingency table which in this context would contain a cases' membership to a particular risk classification by their resolved or not resolved status. The relevant test statistic is estimated at each 'survival' time interval by calculating the expected number of resolutions in each risk group. The null hypothesis is that the survival function for each group is the same. The statistic of focus for the log-rank test is the chi-square statistic, a p-value of less than 0.05 allows rejection of the null hypothesis and suggests there is a statistical significance between groups. The test was performed using the 'survdiff' function under the 'survival' package in R. This gave a statistic of 43.2 on 2 degrees of freedom and a p-value of 0.00000000041 and therefore strong evidence to reject the null and to accept the alternative that there is a statistically significant difference between the risk classifications. A primary disadvantage of this test is that whilst it shows evidence that there is a significant difference between groups in terms of their time to resolution, it does not specify between which risk classifications this difference is. In other words, whether the significant difference is between low and medium risk, medium and high or low and high. This draws attention the earlier mentioned downfall of the Kaplan-Meier estimator in that it can only compare a limited number of groups (Mills, 2011). This will be addressed in later sections by the semi-parametric Cox models, and the findings from the Kaplan-Meier curves will be treated as part of the explanatory analysis.

4.1.2. Investigating Explanatory Variables

Extending the preliminary analysis, Kaplan-Meier curves were plotted for explanatory covariates selected from the data, individual log-rank tests were then performed on each to determine any statistically significant differences between their groups. The full list of explanatory variables used here and in the succeeding analyses is given in Table 8.

Table 8: Explanatory Variables

Variable Name	Description
mfh_risk	Risk classification of missing person
age_cat	Age bracket of missing person
mfh_sex	Sex of missing person
mfh_ethnic_code	Ethnic background of missing person
where_description	Summary of where person stayed whilst missing
return_description	Summary of how missing person was located
likelihood_description	Likelihood of going missing again
mfh_missing_status	Missing status: "Adult Missing from Hospital" "Missing Child/Youth (Under 18 Years) - Cared for by any Local Authority" "Other Missing Adult" "Other Missing Child/Youth (Under 18 Years)
mfh_uk_sirene_category	UK Sirene category

mfh_absconder_id	1 = Juvenile in need of protection or who poses a threat 2 = Adult in need of protection or who poses threat 3 = Adult not in need of protection and not posing a threat Absconder indicator

Individual log-rank tests were conducted for each of the covariates as predictors of time to resolution. Using the null hypothesis that the survival functions were the same for each group within the covariate, the tests found a significant difference between groups for every explanatory covariate. As stated before however, it is not known for the covariates with multiple groups where this significant difference lies between. To better understand where this difference may be, the survival curves and hazard rates for some variables were visually inspected. An example is displayed in Figure 4 which shows the hazard rate separated by 'age_cat'.

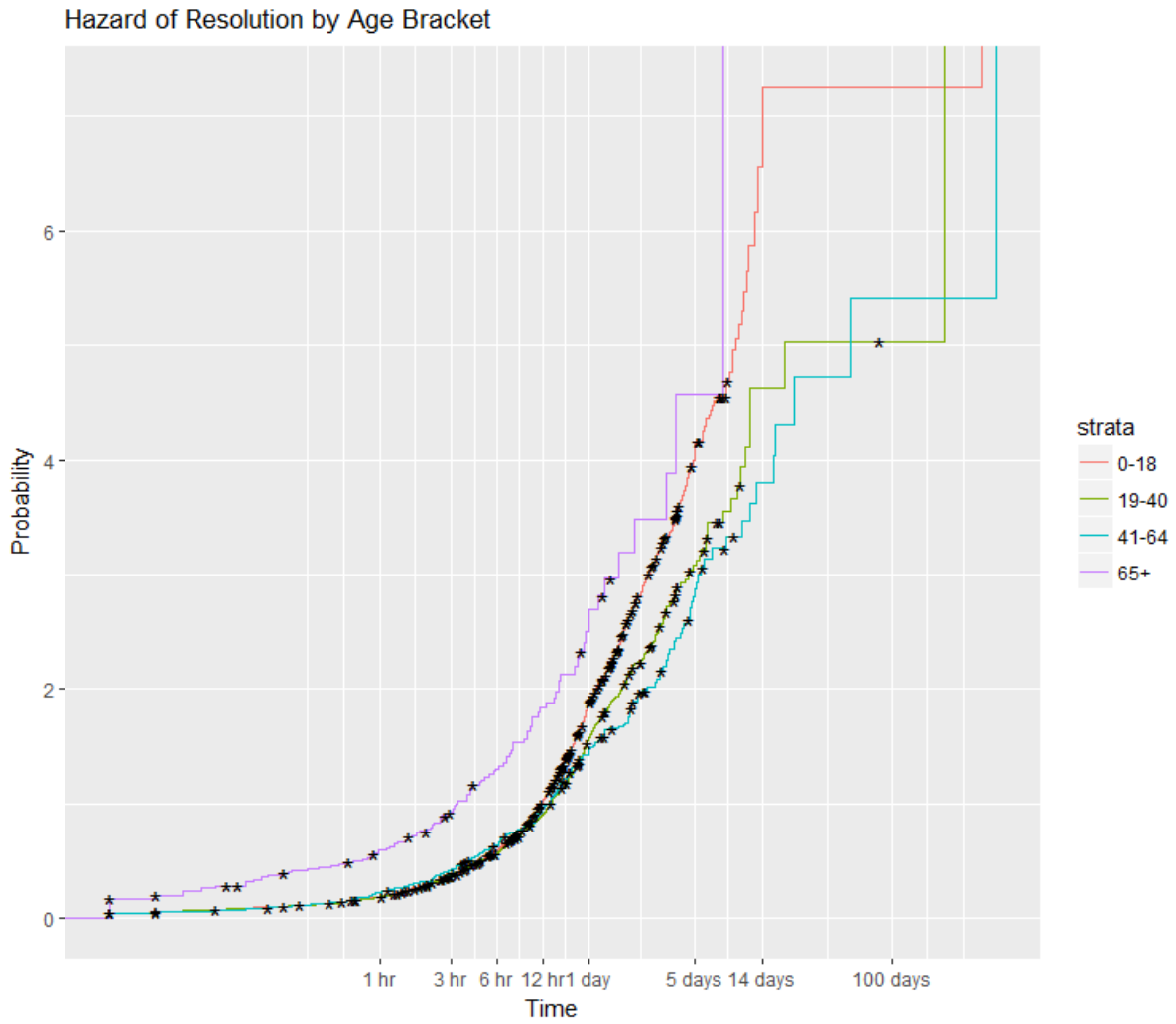

Figure 4: Hazard Rate of Resolution by Age Bracket

Whilst all curves rise and show an increasing likelihood of resolution over time, the rate for missing persons aged 65 and over increases much faster than the other brackets, suggesting that these cases tend to be resolved much quicker. For the remaining brackets, the hazards

appear very close to one another within the first 12 hours of the missing persons report being created and then begin to split after the 1-day period has past, with cases relating to persons between ages 0-18 being resolved sooner than those for the middle age groups. This could relate to the higher risk classifications more likely to be given to the highest and lowest age bracket, with almost half of cases relating to 65 and overs being classified as high risk and over 90% of 0-18 MFH cases being deemed medium risk, as found in the initial data exploration in section 2. Further details on how the multiple covariates and the groups within them relate to the time to resolution will be discussed in succeeding chapters.

4.1.3. Individual Risk Factors and Time to Resolution

Following the previous analysis investigating the individual risk factors that most influence a high risk classification and the present analysis into how risk classifications influence time to resolution, it was of interest to examine how the individual risk factors affect time to resolution. Based on the classification results, it was believed that those risk factors significant in classifying a high-risk missing person would also have a significant influence on resolution time. To determine significance, a log-rank test was performed on each of the 19 risk factors individually, testing the null hypothesis that the survival functions of both the group that replied 'Yes' to having that risk and the group that stated 'No' were the same.

The risk factors that were found to have a significant difference between groups with and without the risk based on the chi-square statistic differed somewhat to the risk factors found to be significant in their influence on risk classification in the earlier logistic regression. Found not to be significant were the risk factors relating to behaviour that is out of character, suspicion that the missing person is subject to a significant crime, failure of the missing person to complete their intentions, previous harm to the missing person whilst missing, if there is any essential medication that they may be missing, if they have suffered ongoing bullying or harassment, whether they suffered violence or racism prior to their disappearance and whether they have any employment, education or financial problems. The remaining risk factors were all found to be significant at the 5% level, plots of the Kaplan-Meier curves indicated whether having the risk factor was associated with a shorter or longer resolution time. Risk factor #1 which relates to the missing person being vulnerable due to age or other infirmity was found to be highly significant with the plotted curves suggesting that cases relating to those deemed vulnerable are resolved faster than those which are not. Missing persons that have indicated they are likely to commit suicide were also found to be significant on resolution time with those that have indicated suicide also having a faster time to resolution. The next significant risk factor related to whether there was a reason for the person to go missing, a plot of the curves showed the survival functions to be very close together, though it appeared that cases where there was a risk of a reason to go missing were not resolved as soon as those without a reason. Risk factor #6 relating to cases with indications that the person had made preparation for absence suggested that those whom had not prepared prior to being missing were resolved sooner than those who had. Risk factor #8 referring to if the missing person had relationship problems or a recent history of conflict or abuse found cases without this risk more likely to be resolved quicker, similarly risk factor #9 relating to the missing person being a victim or perpetrator of domestic violence found missing persons who were believed to be neither of these things more likely to have a shorter resolution time. The three remaining significant risk factors were any physical or mental health problems, if the person is on the Child Protection Register and whether it was felt the

person was unable to safely interact with others or an unknown environment. Plots of the survival function suggested that cases with any of these risks were predicted to have a shorter resolution time than if these risks were not present.

Whilst the earlier analysis showed the significant difference between each of the three risk classifications and their time to resolution, a look at individual risk factors provided a closer view of the types of cases that take the longest time to resolve.

4.2. Summary of Kaplan-Meier Estimator

The Kaplan-Meier estimates provide a starting point for understanding the relationship between the initial risk classification given to a missing person and the time it takes to resolve that case based on a 1 ½ - 2 ½ year follow up time and the perception of resolution as the location of the missing person being known and the outcome being recorded. The estimates do not indicate which groups the significant difference is between, nor do they provide the size of the effect that the variables have on time to resolution. These drawbacks are assessed in later sections.

Examining the influence of the individual risk factors in the latter part of this section highlighted in more detail the types of cases that take longer to resolve, though the analysis only provides a value of significance and not the reasons for these cases having a longer resolution time and so conclusions are speculative. It may be that these cases, for instance those where there is indication of preparation for absence and so a longer resolution time are more complex and are therefore more difficult to resolve. It could be that the risk factors that take longer to resolve though are not considered important in high-risk assignment such as the missing person having relationship problems or a recent history of abuse require a more intense police response than previously given, or increased consideration in higher risk assignment. Whilst this could lead to further discussion into the risk factors that should be classed higher in the risk assessment, the succeeding analyses focus in more detail on how the explanatory covariates shown in Table 7 affect time to resolution.

5. Cox Proportional Hazards Model

As briefly introduced in Section 4, the Cox Proportional Hazards is a semi-parametric model used in event history analysis, as oppose to the preceding non-parametric Kaplan-Meier estimator. Lewis-Beck et al (2004) state that in studies of human behaviour such as this, the Cox model is often preferred due to being less restrictive than non-parametric models. Semi-parametric make no assumption on the distribution of the duration, or in this case resolution times. A key concept as with most models is the hazard rate; the rate at which ‘survival’ time ends or the probability that the subject will experience the event of interest at a particular time (Lewis-Beck et al, 2004), the hazard of experiencing the event is conditional on its history. For Cox models, of focus is usually what is referred to as the ‘hazard ratio’ rather than the hazard rate, as will be explained in further detail shortly.

Mills (2011) discusses further advantages of the Cox model, stating that whilst more flexible than non-parametric it is also a robust model in that it generally fits the data well. Furthermore, it allows for the inclusion of time-varying covariates, though the present analysis focusses on a model of fixed covariates as the variable values remain fixed regardless of time. Mills (2011: 87) expresses the Cox model of fixed covariates in scalar form as:

$$h_i(t) = h_0(t)\{\exp(\beta_1 x_{i1} + \cdots + \beta_k x_{ik})\}$$

In the expression $h_i(t)$ represents the hazard rate for individual *i* at time *t*, shown as a function of two factors. Cox models do not have an intercept term and so $h_0(t)$ refers to the baseline hazard function which is understood as the hazard for a subject with all covariate values of zero. The expression can be written in a linear form of the log-hazard or as a multiplicative form of the hazard.

Whilst the Cox model makes no assumption on the shape of the hazards, its key assumption is that the hazards are proportional (Flynn, 2012), as shown by the full name of the model. Mills (2011) states the reason for this name is because in the model the hazard for any subject is a fixed proportion of the hazard for any other subject. Proportional hazards can be understood as the ratio between the hazards of two or more individuals or groups remaining constant throughout time; referred to as the hazard ratio. This proportionality can be viewed through plots of the log-hazards in which they should appear parallel to one another. It is the hazard ratio rather than the hazard rate that is easier to interpret from the fitted Cox model and is therefore often the value of focus in Cox analysis. The key difference between the rate and the ratio is that the hazard rate is the probability of experiencing the event in a certain time interval given that the subject did experience the event in all the prior time intervals, the hazard ratio estimates the ratio between the hazard rate of one group, for instance the medium-risk group and another group for instance the high-risk group.

Mills (2011) goes on to explain that for the Cox model, parameters are estimated through partial likelihood, as introduced in Cox (1972). The likelihood function is partial as it only describes probabilities for the subjects who experience the event of interest and ignores the ones that do not, the censored. Maximum likelihood on the other hand which is used for most statistical models explains the joint probability of obtaining the observed data for all subjects in the data as a function of the parameters in the considered model.

5.1. Fitting the Model

The Cox models were fitted using the 'coxph' function in the 'survival' package. The hazard ratio was modelled as predicted by the explanatory variables; the full list of which is presented in Table 7 in Section 4. The chosen method for dealing with potential tied survival times was the Efron method (Efron, 1977), which is the default method in R and R Studio. Alternative methods include the Breslow approximation which is generally not recommended and the exact-likelihood method which is computationally slower and does not produce much more accurate estimates (Lewis-Beck, 2004).

Due to the large number of variables and levels within the variables, a forward selection procedure of variable inclusion was chosen. The first model thus fitted was the hazard rate as predicted by only risk level, 'mfh_risk'. The rate was plotted on a log-transformation of the time variable, 'days_missing'. A summary of the first model found that the risk classification of the subject was a significant term in the model and so it remained. The second variable included was the subject age bracket, 'age_cat'. To test if either of the variables were not significant terms in the model and so needed to be removed, a chi-squared test was performed using the 'drop1' function. It was found that both terms were highly significant and so remained. This procedure of including one additional explanatory variable and testing significance of the new model with a chi-square test and removing terms that were non-

significant was repeated until a main-effects model of only significant terms remained. This required eight steps and left a model of eight significant terms: ‘mfh_risk’, ‘age_cat’, ‘mfh_sex’, ‘where_description’, ‘return_description’, ‘likelihood_description’, ‘mfh_missing_status’ and ‘mfh_uk_sirene_category’.

The adjusted curve was plotted for a subject of average covariates in the main effects model, displayed in Figure 5. Unlike the Kaplan-Meier curves, the adjusted curves for the Cox model were plotted using the ‘survminer’ library package (Kassambara, 2017). The time variable along the x axis has again been plotted in log form to focus on the main area of the curve and the time intervals manually added as tick marks.

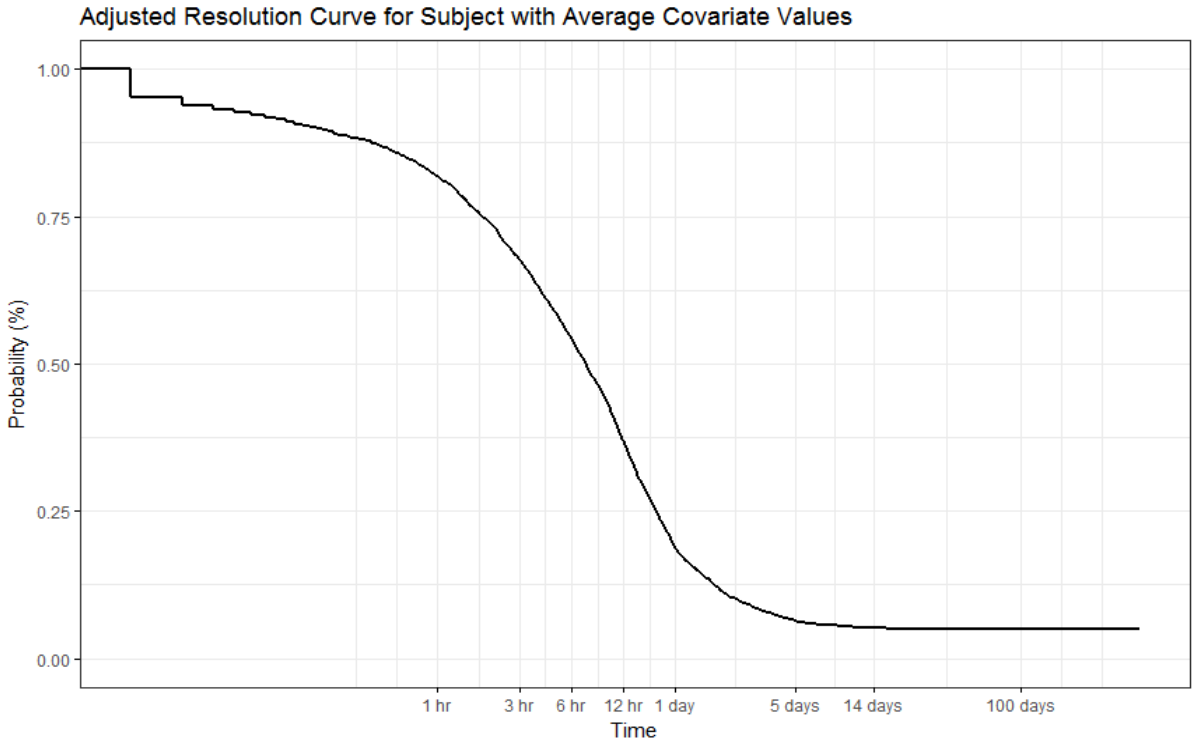

Figure 5: Probability of No Resolution for Main Effects Cox Model

The curve suggests that 50% of cases reach resolution in the first 6-9 hours of the report being created, based on a subject with average covariate values of the significant variables in the main effects model, 25% of cases remain unresolved at around 12-18 hours. Unlike the survival curve produced by the Kaplan-Meier estimator, the Cox curve suggests a much slower decline in cases reaching resolution after the 1-day period, with a much flatter curve between the 1-5 day time intervals. The curve was plotted again, though this time separated by risk level to see the influence of this covariate on the hazard ratio.

Figure 6 shows that based on the main effects model, high-risk cases are more likely to reach resolution than the two other risk levels in the first 24 hours of the report being created, 75% were predicted to reach resolution between the first 12-18 hours. The hazard rates for high-risk and medium-risk then appear to cross between the 1-2 ½ day tick marks with medium-risk cases having a steeper curve and a predicted faster time to resolution than high and standard risk cases. Standard cases were the least likely to reach resolution throughout all time intervals.

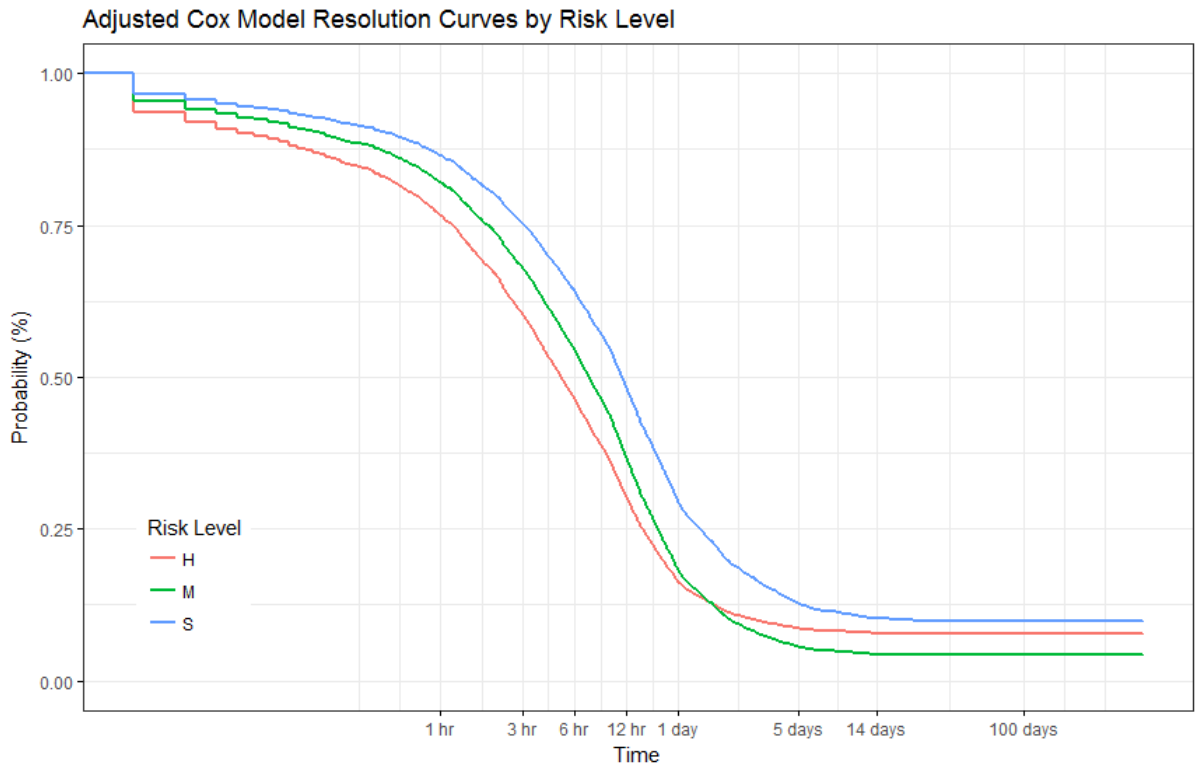

Figure 6: Adjusted Curves for Main Effects Cox Model Separated by Risk Level

The estimates from the main effects model are presented in Table 9 alongside the exponential estimates which are used for interpretation and the exponentials of the 95% confidence intervals.

Table 9: Hazard Ratio Estimates from Main Effects Cox Model

			95% Confidence Interval	
Variable Description and Reference Category	**Estimate**	**Exponent of Estimate**	**Lower Bound**	**Upper Bound**
Risk Level: Medium	-0.16	0.85	0.77	0.95
Risk Level: Standard	-0.31	0.73	0.62	0.86
Age Category: 19-40	-0.02	0.98	0.71	1.34
Age Category: 41-64	-0.04	0.96	0.69	1.34
Age Category: 65+	0.70	2.02	1.42	2.88
Sex: Male	-0.11	0.90	0.84	0.95
Where Stayed: Hotel/Similar	-0.35	0.70	0.46	1.06
Where Stayed: Sexual Exploitation	1.15	3.16	1.49	6.69
Where Stayed: With another Missing Person	0.20	1.23	0.82	1.83
Where Stayed: Unknown	0.40	1.44	1.00	2.08
Where Stayed: Refuge/Similar	-0.90	0.42	0.23	0.76
Where Stayed: Slept Rough	-0.39	0.68	0.46	1.00
Where Stayed: Place Previously Known	0.39	1.49	0.94	2.34

Where Stayed: No Known Connections	-0.14	0.87	0.54	1.39
Where Stayed: With Friend	-0.01	0.99	0.69	1.44
Where Stayed: With Partner/ Ex Partner	-0.18	0.84	0.55	1.28
Where Stayed: With Person Just Met	-0.15	0.86	0.52	1.41
Where Stayed: Person/Place from Original Information	0.24	1.28	0.74	2.19
Where Stayed: Travelled Abroad	-1.21	0.30	0.15	0.60
Where Stayed: With Other Relative	-2.00	0.82	0.56	1.20
Returned: Found Deceased	-0.85	0.43	0.24	0.77
Returned: Found by Family/Carer	0.56	1.76	1.37	2.24
Returned: Found Harboured/Abducted	-0.38	0.69	0.17	2.81
Returned: Found in Hospital	0.36	1.44	0.97	2.14
Returned: Found by Police	0.55	1.73	1.37	2.17
Returned: Unknown/Other	-5.56	<0.005	<0.0006	0.03
Returned: Own Accord	0.57	1.76	1.40	2.22
Likelihood of Missing Again: Unlikely	0.09	1.09	0.98	1.22
Likelihood of Missing Again: Very Likely	-0.20	0.82	0.76	0.87
Missing Status: Missing Youth in Care	0.16	1.17	0.82	1.68
Missing Status: Other Missing Adult	0.02	1.02	0.86	1.21
Missing Status: Other Missing Child	0.33	1.40	0.97	2.01
UK Sirene Category: 2. Adult in need of protection or who poses threat	0.18	1.19	1.06	1.34
UK Sirene Category: 3. Adult not in need of protection and not posing a threat.	NA	NA	NA	NA

The ratios displayed in the second column are multiplicative in their effect on the hazard, a value of above 1 indicates that the variable is associated with an increased hazard of case resolution whilst a value less than 1 shows a decreased hazard of the case being resolved. The reference categories for each level of the variable are stated and so interpretation understands the estimate of the level in comparison to the one that is not stated, for example each age category represents effects compared to if the subject was between ages 0-18. With regards to risk classification, it can be seen immediately that high-risk cases are the most likely to be resolved, with the hazard of resolution for medium-risk cases being 85% and standard-risk

being 73% of that for high-risk cases. Missing persons aged 65 years and over are the age group most likely to be resolved quickly with twice the hazard of resolution than persons aged between 0-18 years, those in the middle age brackets had the lowest odds of being resolved when all other covariates were held constant. Regarding sex, females were most likely to be resolved with males having a resolution hazard that was 90% of that for females.

The reference category for where the person stayed whilst missing is in hospital or policy custody and so all estimates are in comparison to those cases. The least likely to be found were those staying in a hotel or similar premises, with a hazard of 0.30 indicating these cases were 70% less likely to be resolved than those that stayed in hospital or police custody. The most likely to be resolved were those involved in sexual exploitation who were over three times more likely to be found than those in the reference category. The reference for how the person was returned is if that person was arrested, the most likely to be resolved were therefore if the person returned of their own accord or if they were found by family with the odds of resolution increasing by 76% compared to if they were arrested. Those found by police followed this with the hazard increasing by 73% compared to if arrested. For the likelihood of going missing again, it was those who were deemed likely to be missing again that had a higher hazard of resolution, with those deemed very likely to be missing again having the lowest. The reference for missing status was missing adult in hospital, the reference had the lowest odds of being found and other missing children not under local authority care had the highest, to be a child not under care increased the odds of being located by 40% in comparison to if the person was a missing adult from hospital. Finally, the reference for UK Sirene category is a juvenile in need of protection or who poses a threat. Estimates were not provided for adults not in need of protection or posing a threat but adults who needed protection or posed a threat were 19% more likely to be located than juveniles. Several diagnostic tests were then conducted on the model to check goodness-of-fit and test the proportional hazards assumption, further details are given in the following subsection.

5.2. Diagnosing the Model

The first test was to check the goodness-of-fit for the main effects model. Using some summary statistics from the model, the likelihood ratio test comparing the main effects to the null model gave a deviance of 1275 on 33 degrees of freedom and a p-value of 0 suggesting very significant evidence to reject the null hypothesis that the coefficients are equal to zero and accepting that they do have a significant effect on time to resolution. The Wald test gave a deviance of 528.3 on 33 degrees of freedom and a p-value of zero, also suggesting significant evidence to reject the null that the coefficient is equal to zero and the main effects is the better fitting. The overall goodness-of-fit was then visually inspected through a plot of the Cox-Snell residuals. A well-fitting model would have produced a straight line through residuals at a 45-degree angle (Mills, 2011), which was not the case.

A key assumption as stated earlier is that the hazards in a Cox model are proportional. This assumption was tested using the estimates from Schoenfeld residuals. These residuals are the observed minus the expected values of the covariates at each ‘failure’ or resolution time (Mills, 2011). Due to the large number of variables and levels within those variables, the residuals were understood through their estimates as oppose to visually through a plot. The p-values of the estimates for each covariate and the global model were used to test the null hypothesis that the hazards are proportional. A p-value of less than 0.05 would thus reject this

hypothesis and the model would fail the assumption. The proportional hazards test gave a global p-value of zero, suggesting that the hazards were not proportional.

5.2.1. Stratification

Mills (2011) states that there are two main ways of dealing with non-proportional hazards such as this, the first is to include interactions into the model and the second is stratification. Several combinations of interactions were added to the model and the assumption re-tested, though all returned with p-values of less than 0.05. It was therefore chosen to stratify some of the covariates. Stratifying variables allows them to have their own baseline hazard function and assume the coefficients are constant across the different strata. The most advantageous way to stratify is to choose the variables that are not of primary interest to the analysis as estimates are not provided for the stratum (Mills, 2011). Stratification was firstly applied to the likelihood of going missing again variable as this was not of focus, though the p-value did not increase. A p-value of 0.001 was achieved when stratified by subject sex, where description, return description, age category and likelihood description, though this still did not satisfy the proportional hazards assumption. The final model that did fulfil the proportional hazards assumption with a p-value of 0.971 that could not reject the null was that which was stratified by risk level, sex, where description, return description and likelihood description. This unfortunately meant that no estimates were given for the risk level variable of primary interest.

Before interpreting the estimates, further diagnostics were performed to the stratified model. The likelihood ratio test gave deviance of 51.3 on 7 degrees of freedom and a p-value of < 0.001 suggesting strong evidence of significance. It was known that the model satisfied the proportional hazards assumption and so the final test was for nonlinearity. This was assessed using Martingale residuals plotted against covariates to form component-plus-residual plots (Mills, 2011). The plot is shown in Figure 7.

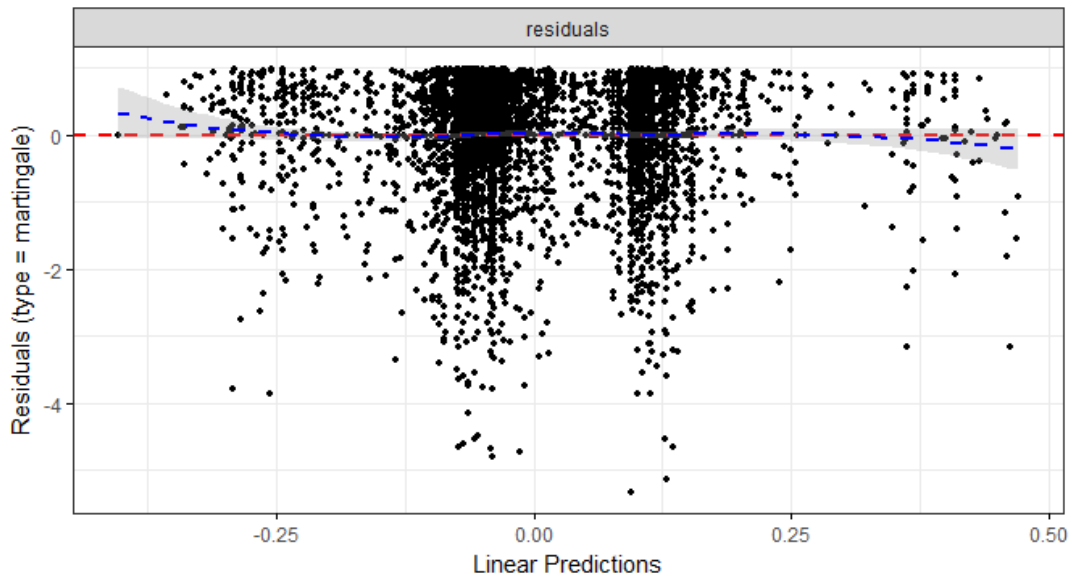

Figure 7: Martingale Residuals for Stratified Cox Model

The dashed red line represents the zero value, the blue line is relatively flat and does not deviate much from the zero line which suggests no changes to the functional form are needed.

The coefficients from the stratified main effects model were therefore used for interpretation, the estimates are presented alongside their exponentials and exponentials of confidence intervals in Table 10. The exponential of the estimate represents the multiplicative effects on the hazard rate, known as the hazard ratio. Estimates are not given for stratified variables.

Table 10: Hazard Ratio Estimates for Stratified Cox Model

			95% Confidence Interval	
Variable Description and Reference Category	**Estimate**	**Exponent of Estimate**	**Lower Bound**	**Upper Bound**
Age Category: 19-40	-0.12	0.89	0.62	1.26
Age Category: 41-64	-0.09	0.91	0.63	1.31
Age Category: 65+	0.38	1.46	0.97	2.18
Missing Status: Missing Youth in Care	0.12	1.13	0.75	1.70
Missing Status: Other Missing Adult	0.01	1.01	0.83	1.23
Missing Status: Other Missing Child	0.29	1.33	0.88	2.01
UK Sirene Category: 2. Adult in need of protection or who poses threat	0.20	1.23	1.06	1.42
UK Sirene Category: 3. Adult not in need of protection and not posing a threat	NA	NA	NA	NA

The ratios in the second column are interpreted in the same way as in Table 8 for the main effects model. For instance, to be in the 19-40 age bracket multiplies the hazard of resolution by 0.89 compared to being between ages 0-18, or in other words the hazard for 19-40 is 89% of that for 0-18s. The 41-64 bracket is also associated with a reduced hazard, whilst being aged 65 or above increases the hazard of resolution by 46% compared to being between 0-18 years old. The missing status rows are in comparison to if the subject was an adult from hospital, the biggest difference is if the missing person is any other missing child not under local authority care in which the hazard of resolution increases by 33% compared to if they were an adult missing from hospital. The first UK Sirene category which is not displayed in the table relates to juveniles in need of protection or pose a threat of which all juveniles (below age 18) are classified. To be an adult in need of protection is to increase the hazard of the case being resolved by 23%. In summary, the greatest 'hazard' or likelihood of a case being resolved applies to juveniles between ages 0-18 who are not under local authority care. The benefits of the Cox model is that effect sizes are estimated for each group in multiple variables. What remains unknown are the reasons behind the hazards, it may be that cases involving young people are more likely to be classed as high risk and therefore receive more police attention and are thus resolved quicker, it could be that these cases are simpler to resolve than a missing adult who has more resource to extend their missing period such as money and modes of transport. Implications from the results will be discussed in further detail in the final section.

5.3. Summary of Cox Model

The Cox model improves on the Kaplan-Meier estimates as it provides a full main effects model outlining the most important factors in predicting the time to resolution of a missing from home case based on the data available. Additionally, it provides estimate of the effect size for each group within each variable, for instance in the stratified model it can be understood that those between ages 19-40 are the least likely to reach resolution of the age groups and adults that are missing from home have the lowest chance of resolution based on missing status. This information provides a basis into understanding cases that are more complex in solving or receive less police response than others and should aid police targeting.

A disadvantage to the analysis is that the final stratified model, whilst fulfilling the proportional hazards assumption that is key to the Cox model, does not provide estimates for some of the primary factors of interest. It cannot be understood from the stratified model for instance the value of the difference in being a high-risk case as oppose to a standard in terms of predicted time to resolution. It could be said from the analysis that whilst the stratified model may be more accurate and diagnostically correct, it is the main effects model that is perhaps more useful for understanding missing person resolution time and informing further work. In addition to the effects of each category group, the adjusted survival curves produced from the main effects model that are presented in Figure 5 and Figure 6 highlight the key time intervals that see the most cases reaching resolution. In Figure 5 it could be interpreted that the first nine hours of a missing person report being created are the most crucial in successfully locating a case, with 50% of reports reaching resolution in this time. A further 25% are resolved by the first 18 hours of report creation. The flattening curve past the 5-day tick mark suggests that cases that remain unresolved at this time point are less likely to reach resolution. Of concern in Figure 6 is the crossing of survival functions for high and medium-risk missing persons after the 1-day period, in which high-risk cases are no longer the most likely to reach resolution. It could be assumed that the large volume of medium-risk cases after this point consume police resource that cannot then be distributed to the more complex high-risk cases that are not resolved in the first 24 hours. This may raise questions as to the assignment of the medium-risk classification and the revision of assessment criteria so that cases continuously classified as medium-risk such as those relating to children aged 18 and below be reconsidered as standard or high-risk depending on the nature of the report to allow better dissemination of police resource to cases that are more in need of police action.

Though more advanced than the non-parametric estimator, the Cox model is also subject to disadvantages. Of focus for this analysis is the assumption by both the Kaplan-Meier estimator and the Cox model that all events are independent of one another, despite the research knowing that many of the missing person cases relate to the same people going repeatedly missing. The following section will introduce ‘frailty models’, an extension to event history analysis for handling recurrent events such as repeatedly missing persons.

6. Frailty Model

The previous event history or ‘survival’ analyses that fitted Kaplan-Meier estimates and Cox Proportional Hazard models both assumed that all subjects, that is missing persons, were homogenous and the events, in this context missing cases being resolved, were considered independent of each other. A drawback of this assumption is that it does not account for subjects that may be more prone to becoming missing person cases that are

subsequently resolved, termed in this methodology as more 'frail'. A more complex extension to event history analysis that accounts for this is the frailty model which involves correlated survival data. Broström (2012) states that frailty models in survival analysis correspond to multilevel or hierarchal models in linear or generalised linear regression. Correlation of event times can occur when a subject experiences an event more than once (Mills, 2011), for instance if the same person is missing and located more than once in the data follow-up period.

Mills (2011) refers to frailty as an unobserved random proportionality that alters the hazard function of a subject or related subject. There are several types of frailty including shared, unshared, nested, joint and additive. Of interest in the MFH data is shared frailty which can apply to both recurrent events and to clustering in groups. Therneau and Grambsch (2001) state that individuals within a dataset have different frailties, and those that are the most 'frail' are those more likely to experience the event of interest earlier. They state that whilst it is known in research that individuals are dissimilar and have what is often referred to as unobserved heterogeneity, this variance is often ignored as nuisance. If this unobserved heterogeneity is ignored, parameter estimates may be inconsistent, standard errors may be incorrect and estimates of duration dependency may be misleading (Mills, 2011). Therneau and Grambsch (2001) do state however that the inclusion of this frailty which is also referred to as the random effects of a model has increased in survival analysis over recent years. Therneau and Grambsch state a simple random effects model in survival analysis is the shared frailty model, the computation of which can be understood as a penalised Cox model. Rondeau et al (2012) describe the shared frailty model with a gamma distribution as the most often used for data with recurrent events such as the MFH data which sees the same subjects repeatedly being reported missing and subsequently located. Rondeau et al (2012: 3) express the hazard function of this shared frailty model as:

$$\lambda_{ij}(t|v_i) = v_i\lambda_0(t)\exp\left(\beta^\tau X_{ij}\right) = v_i\lambda_{ij(t)}$$

With the j-th individual $(j = 1, \dots, n_i)$ for the i-th group $(i = 1, \dots, G)$, $\lambda_0(t)$ is the baseline hazard function, X_{ij} is the covariate vector associated with the vector of regression parameters β and the random effect v_i is that which is associated with the i-th group. The random effects are assumed to be independent and identically distributed from a gamma distribution with $E(v_i) = 1$ and $Var(v_i) = \theta$.

6.1. Including Frailty to the Main Effects Model

Frailty was added to the model using the 'frailty' function under the 'survival' package (Therneau and Lumley, 2017). The frailty term was added to the 'id_num' variable which represented the ID variable created at the beginning of the project which assigned each different individual a unique number to allow tracking of repeatedly missing persons. Following on from the Cox analysis in the preceding section, it was summarised that the main effects model of significant terms was more useful due to its estimates for all covariates of interest, though admittedly violating the proportional hazards assumption that is key to the Cox model. The stratified model fulfilled this assumption though did not provide estimates for some of the main covariates of interest such as subject risk classification. It was thus chosen to add a frailty term to both the main effects model and to the stratified model and to interpret both outputs, acknowledging the benefits and limitations of each.

The correlation term was firstly included into the main effects Cox model alongside the significant covariates. The output from the model provides estimates for the coefficients as would a standard Cox model though also provides information on the significance of the frailty term and the amount of variance of the random effects. The model output suggested that the frailty term was very significant with a p-value of less than 0.000000001, thus improving the model that did not account for event correlation. The variance of the random effects was 0.234; the closer this figure is to zero suggests lesser evidence of frailty. The likelihood ratio test gave deviance of 2122 on 701.5 degrees of freedom and a p-value of 0 suggesting the model was a good fit to the data. As explained by Mills (2011), the coefficient estimates can be interpreted the same as in other Cox models, though they are now understood to be conditional on frailty. The exponents of the coefficients were produced to enable interpretation of the hazard ratio, the estimates are presented in Table 11.

Table 11: Hazard Ratio Estimates for Main Effects Model with Frailty

			95% Confidence Interval	
Variable Description and Reference Category	**Estimate**	**Exponent of Estimate**	**Lower Bound**	**Upper Bound**
Risk Level: Medium	-0.18	0.83	0.73	0.95
Risk Level: Standard	-0.37	0.69	0.57	0.84
Age Category: 19-40	0.05	1.06	0.72	1.54
Age Category: 41-64	0.05	1.06	0.71	1.56
Age Category: 65+	0.94	2.60	1.66	3.93
Sex: Male	-0.07	0.93	0.85	1.01
Where Stayed: Hotel/Similar	-0.60	0.60	0.34	0.88
Where Stayed: Sexual Exploitation	1.10	3.02	1.31	6.92
Where Stayed: With another Missing Person	0.03	1.03	0.66	1.61
Where Stayed: Unknown	0.22	1.25	0.83	1.88
Where Stayed: Refuge/Similar	-1.31	0.27	0.13	0.56
Where Stayed: Slept Rough	-0.67	0.51	0.33	0.80
Where Stayed: Place Previously Known	0.29	1.34	0.80	2.25
Where Stayed: No Known Connections	-0.31	0.74	0.42	1.28
Where Stayed: With Friend	-0.18	0.84	0.55	1.27
Where Stayed: With Partner/ Ex Partner	-0.37	0.69	0.43	1.12
Where Stayed: With Person Just Met	-0.38	0.69	0.38	1.22
Where Stayed: Person/Place from Original Information	0.31	1.36	0.72	2.58
Where Stayed: Travelled Abroad	-1.79	0.17	0.07	0.38
Where Stayed: With Other Relative	-0.39	0.68	0.44	1.04
Returned: Found Deceased	-1.06	0.35	0.16	0.74

Returned: Found by Family/Carer	0.47	1.60	1.21	2.13
Returned: Found Harboured/Abducted	-0.69	0.50	0.11	2.26
Returned: Found in Hospital	0.42	1.53	0.96	2.43
Returned: Found by Police	0.52	1.68	1.29	2.20
Returned: Unknown/Other	-5.94	<0.01	<0.01	0.02
Returned: Own Accord	0.51	1.67	1.28	2.18
Likelihood of Missing Again: Unlikely	0.08	1.08	0.94	1.24
Likelihood of Missing Again: Very Likely	-0.20	0.82	0.75	0.89
Missing Status: Missing Youth in Care	0.32	1.37	0.88	2.12
Missing Status: Other Missing Adult	0.02	1.02	0.83	1.25
Missing Status: Other Missing Child	0.47	1.60	1.04	2.48
UK Sirene Category: 2. Adult in need of protection or who poses threat	0.21	1.23	1.06	1.43
UK Sirene Category: 3. Adult not in need of protection and not posing a threat.	NA	NA	NA	NA

The hazard ratio estimates are again given in the second column though are now conditional on frailty, values above 1 suggest an increase in the hazard and those below 1 suggest a decrease in the hazard in comparison to the reference categories. The reference categories are the levels that are not displayed in the table and are the same as those discussed in Table 9 in the previous section. Several of the estimates have changed now that correlation has been accounted for, the differences are greater in some variables than in others. The first difference of interest was in the age categories; in the main effects Cox estimates, to be in the 19-40 or 41-64 age bracket reduced the likelihood of the case being resolved in comparison to if they were between the ages of 0-18. With frailty now accounted for, to be aged 19-40 or to be aged 41-64 increases the hazard of being resolved by 6% compared to if that case referred to a missing person between the ages 0-18. This shows the importance of the frailty term in the model as previous estimates that did not account for this can now be deemed misleading. Also examined were the new estimates for the variable relating to missing status, whilst these stayed in the same direction, the estimates for some levels showed a relatively large difference. The reference category was cases relating to missing adults from hospital, thus to be a missing youth in care increased the hazard by 37% when conditional on frailty, without accounting for frailty the increase was 17%. To be a missing child not under care increased the hazard of resolution by 60%, this as an increase of 40% when frailty was not included.

The curves for the main effects model with and without frailty were plotted for comparison. The curves were plotted using ‘ggfortify’ (Horikoshi, 2017) and ‘ggplot’ (Wickham and Chang, 2016). Again the time axis was plotted in log form to focus on the main area of the curve. The curve without frailty is firstly displayed in Figure 8.

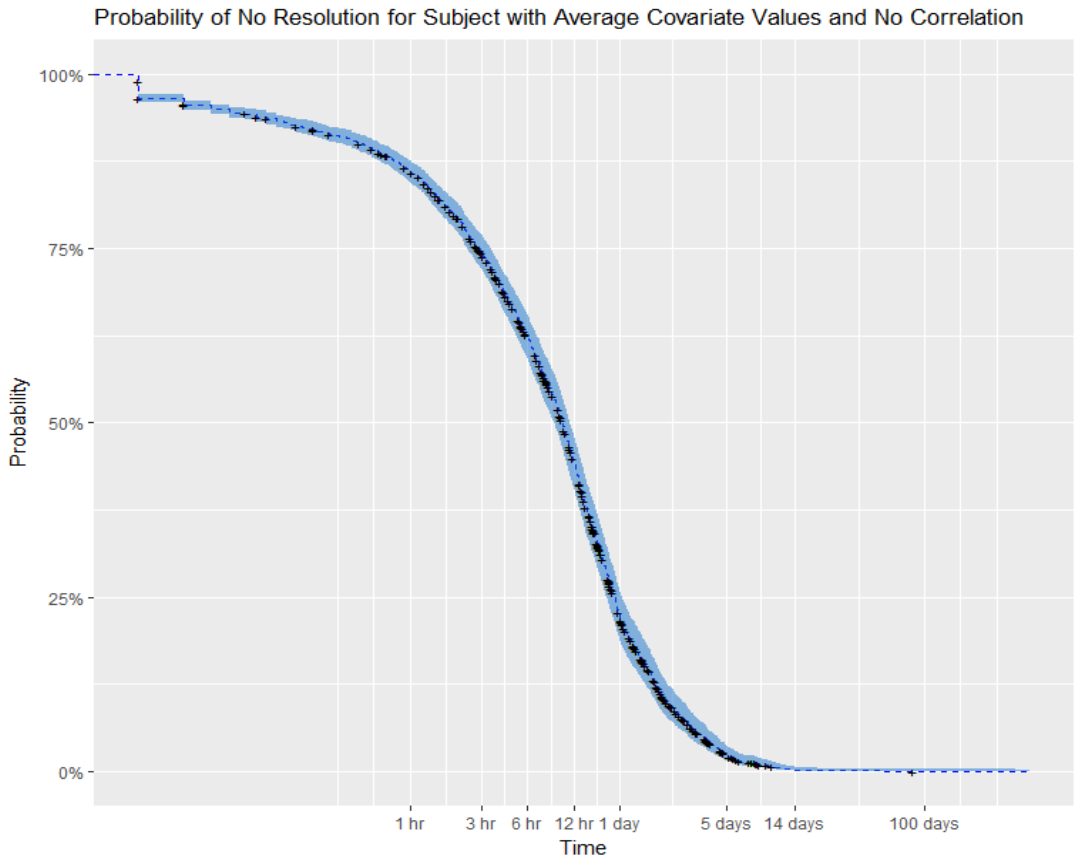

Figure 8: No Resolution Curve of Main Effects Model without Frailty Term

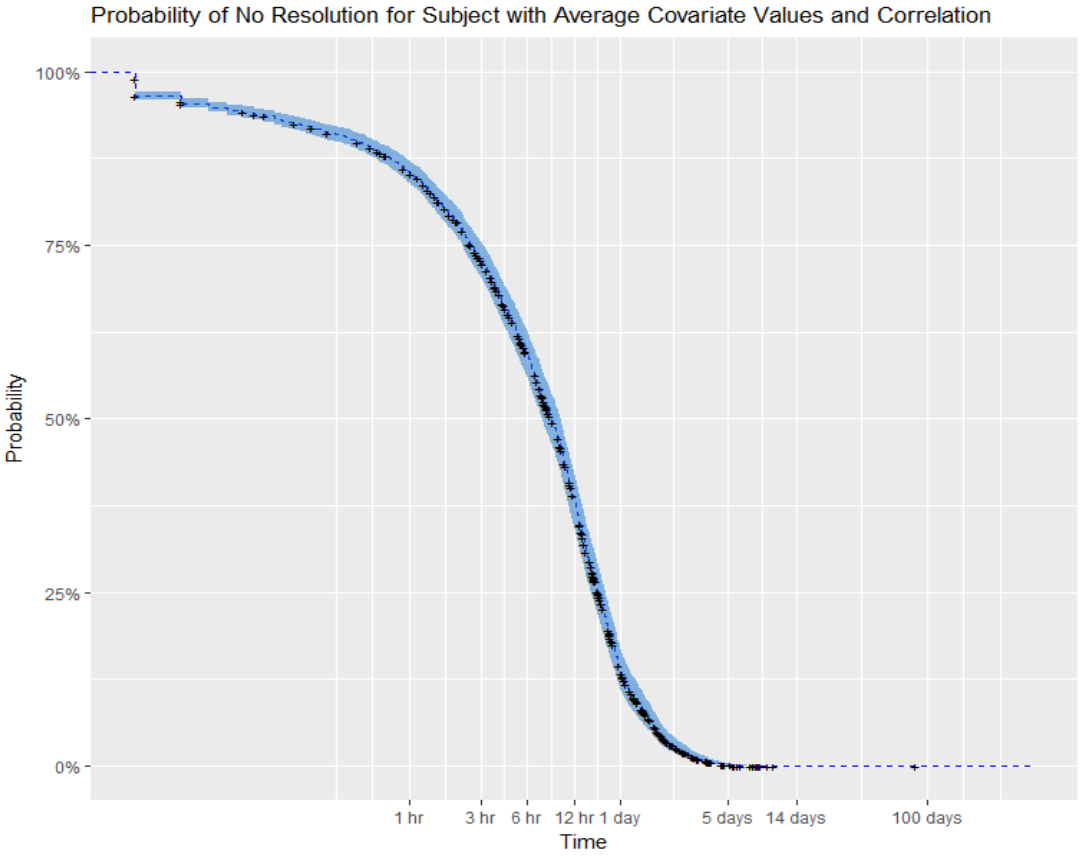

Figure 9: No Resolution Curve of Main Effects Model with Frailty Term

The curve replicates that in Figure 6 in the previous section though with a different graphical package, the shaded area represented the 95% confidence intervals and the asterix symbols represent censored observations. This was then compared to the curve produced for the main effects model that did include the frailty term, displayed in Figure 9.

At first glance the two curves appear very similar, a closer examination can see that the curve for the model that accounts for frailty is much steeper. Without accounting for this correlation, the main effects model suggested 75% of cases would reach resolution within the first 24 hours. When this correlation is modelled, the curve predicts that 75% of cases were actually resolved within the first 18 hours of being created. The curve is also much steeper between the 1-5 day time intervals. This difference shows the benefit of fitting the model with frailty accounted for with a more accurate understanding of the time to resolution and the curve suggesting missing from home cases are resolved faster than initially thought.

6.2. Including Frailty to the Stratified Model

A frailty term was then added to the stratified main effects model which fulfilled the proportional hazards assumption. This model can be deemed more accurate though less useful for future implications due to the estimates of several explanatory variables not being provided. The output from the new model suggested that the frailty term was very significant with a p-value of less than 0.00000001 and so an improved model in comparison to that without frailty. The variance of random effects was 0.065 and so less evidence of correlation than in the main effects model, though this may be expected with some of the variance being accounted for by the stratification. The likelihood ratio test for the model gave deviance of 588.8 on 196.7 degrees of freedom and a p-value of 0 suggesting it was a good fit to the data. The estimates for the new model are given in Table 12 and are interpreted the same as in Table 11.

Table 12: Hazard Ratio Estimates for Stratified Model with Frailty

			95% Confidence Interval	
Variable Description and Reference Category	**Estimate**	**Exponent of Estimate**	**Lower Bound**	**Upper Bound**
Age Category: 19-40	-0.11	0.90	0.62	1.31
Age Category: 41-64	-0.08	0.93	0.63	1.36
Age Category: 65+	0.44	1.55	1.01	2.37
Missing Status: Missing Youth in Care	0.17	1.18	0.77	1.82
Missing Status: Other Missing Adult	<0.01	1.00	0.81	1.23
Missing Status: Other Missing Child	0.32	1.38	0.89	2.12
UK Sirene Category: 2. Adult in need of protection or who poses threat	0.22	1.24	1.07	1.45
UK Sirene Category: 3. Adult not in need of protection and not posing a threat	NA	NA	NA	NA

The estimates of focus are again those in the second column which represent the hazard ratio.

The estimates have changed though do not differ too greatly from those that did not account for frailty and are presented in Table 10. The largest difference relates to the age category variable and specifically the 65+ bracket. The estimate above suggests that to be aged 65 or over increases the hazard of resolution by 55% than if the missing person was in the reference category 0-18, without frailty this increase was 46% and so a change of 9% is the largest difference. In comparison to the main effects model that did account for frailty in Table 11, it is interesting to see that the estimates for age category have again changed direction and the above values suggest that to be in the 19-40 or 41-64 brackets actually decrease the hazard in comparison to 0-18 whilst the main effects model suggest these increase. Based on fulfilling the proportional hazards assumption and also accounting for unobserved heterogeneity, it could be said that the stratified main effects model with a frailty term is that which should be taken as most accurate. However, it is acknowledged that this may not be of the most use for future missing from home recommendations and there are arguments to suggest the main effects model is the more appropriate, as will be discussed in the following summary.

6.3. Summary of Frailty Model

The results from both the main effects and the stratified model demonstrate the benefit of including the frailty term due to its significance and thus improvement to the model. This applies to both the main effects and to the stratified models, the estimates are now more accurate and are understood to be conditional on frailty. With regards to which model estimates should be used going forward, whilst the stratified model may be deemed a better fitting Cox model due to its fulfilment of the proportional hazards assumption, the main effects model without stratification provides more estimates and more importantly it provides estimates for the main variables of interest such as the risk level. Additionally, the frailty model has been deemed as an extension to the Cox model to account for non-proportional hazards. Perperoglou et al (2006) state that when non-proportional hazards are found in a Cox model, a natural step to take is to investigate time-varying covariates. However, this step is not always appropriate, as it was not for the MFH dataset. They state an extension to the model that could explain the behaviour of the covariates is the Gamma frailty model. For the purpose of this research, the hazard ratio estimates taken from the main effects model with the frailty term displayed in Table 11 will be regarded as final and used as a basis in further discussion.

Overall some key findings from the frailty models with regards to time to resolution is that those classified as high risk have the greatest ‘hazard’ or likelihood of resolution, which is somewhat expected. The age category variable suggests that those in older brackets, in particular those aged 65 or over are more likely to be resolved than those in younger age brackets. Additionally, UK Sirene categories suggest that adults in need or protection or who pose a threat are more likely to be located than juveniles in need of protection or who pose a threat. On the other hand, the missing status category contradicts this in suggesting adults missing from hospital are the least likely to be resolved and other missing children not in the care system have the highest hazard, this may pose an area for further analysis. Regarding where persons stay whilst missing, those involved in sexual exploitation have the highest hazard of resolution whilst those who travelled abroad have the lowest, followed by those who stayed in a refuge or similar premises. Missing persons found by police were the most likely to be resolved, followed by those who returned of their own accord. Disregarding the

unresolved outcome level, those who were found deceased had the lowest hazard of resolution, presumably due to an increased difficulty in finding deceased missing persons.

7. Discussion

The primary objective of this project was to investigate the appropriateness of the current risk classifications of MFH cases based on the time taken for the case to be deemed resolved. Secondary objectives include highlighting the demographics of those reported missing, exploring the way in which 'High Risk' is defined and examining the influence of repeatedly missing persons. The first objective addressed was the demographics of the missing persons in the dataset, which was explored through frequencies and cross-tabulations in SPSS Statistics and outlined in Section 2. Some key findings included that over half of all missing persons were male, 71.6% of all were between 0-18 years and 41.8% of all cases related to missing children under local authority care. The most common areas were Blackpool and Preston. With regards to risk, the large majority were medium risk at 83.1%. Some findings raised questions surrounding risk classification and its appropriateness, for instance 92% of missing children from care were assigned medium risk, though 47.4% returned of their own accord. 89.3% of missing children not in the care system were classified as medium risk and 46.5% returned of their own accord. The only missing category to be found deceased were other missing adults not in hospital. Of all missing person cases, less than 10% were deemed unlikely to be missing again. The classification of 'High Risk' was then investigated through a logistic regression treating the binary 'High Risk' or 'StdMed Risk' variable as dependent and the individual risk factors and explanatory. In summary, it was found that a high risk classification is most likely given to someone who displays a number of features; an individual who is behaving in a way that is out of character, is suspected to be subject to a significant crime, has indicated they are likely to commit suicide, has physical or mental ill-health, is on the Child Protection Register, is believed to be unable to interact safely with others or an unknown environment or has any other unlisted risk factors believed to be important to the level of risk. Factors that lessen the likelihood of being deemed high-risk are if the missing person has financial, employment or education problems or if the person has a drug or alcohol dependency.

The 'appropriateness' of the risk classifications was investigated in various stages, predominantly through event history techniques; the Kaplan-Meier estimator and the Cox Proportional Hazards Model. This particular model was later extended to a frailty model to account for the effects of repeat missing persons. Overall, it was found that risk classification does have a significant influence on the time to resolution with high risk cases more likely to reach resolution, though plots of these curves show that this changes over the time period, with medium risk cases over taking high risk in the speed of resolution after the first 24 hours. It was felt that should cases have been classified under a different risk level such as high or standard dependent on the case, then the hazards would not have crossed and high-risk cases would have remained as being solved quicker and this would show a different shape to the curves. It was felt this crossing could have been due to the vast number of medium-risk cases that the police dealt with which could take resource away from the perhaps more complicated high-risk cases that take longer to resolve. Several other explanatory variables were found to be significant on the time to resolution: age bracket, sex, where stayed whilst missing, how the person was returned, the likelihood of them being missing again, their missing status and the UK Sirene category.

The appropriateness of risk classification was also investigated through estimating time to resolution as predicted by the individual risk factors. It was found that some of those risk factors that are of important influence in high-risk assignment are not of significance in predicting time to resolution. Behaviour that is deemed out of character and suspicion that the missing person is subject to a significant crime both significantly increased the likelihood of a high-risk classification though did not significantly influence time to resolution. This could suggest that greater priority needs to be given to risk factors that indicate longer resolution time when making the initial risk assessment to reduce the number of long cases that require police attention throughout and perhaps less priority given to risk factors that do not affect this time to resolution. This of course only considers risk based on resolution time and does not account for other factors such as harm that comes to the person whilst missing.

7.1. Limitations

As covered in the relevant sections, there are limitations to each analysis method that was used and improvements could be made, though these are discussed elsewhere. The first noted drawback to the project was the large amount of missing data. It was chosen to take a complete-case analysis and remove all observations that contained missing values. The data was missing for several reasons; predominantly due to the information not being available and data input errors. There were particular issues with the input of the date the report was created and the date the person was (or was not) found. Whilst complete-case analysis such as this is a common approach due to its simplicity, it can lead to bias and an insufficient analysis (Horton and Kleinmen, 2007). In addition, the great loss in sample size, which in this case was 5952 to 4746 observations, reduces the statistical power of the analysis (Madden et al, 2016). With fewer time constraints, imputation methods of dealing with the missing data would have been explored to improve the quality of the analysis.

A second limitation to the analysis was the data only provided the risk classification that was given to the missing person at the time of reporting. It is known that this risk classification can change throughout the time-period that the person is missing due to changes in circumstance or new information. The risk level of a missing person may change several times whilst the case is open. It is known from the analysis that the risk level influences time to resolution, and so it would be expected that the changing risk level throughout a person's missing time-period would also influence their time to resolution. This is an area of interest for further investigation.

7.2. Extensions

As discussed throughout, there are several areas left open for further research. Though the main extension of interest in this project would be to investigate the 'resolution time' based on the several different outcomes that a case could have as opposed to focussing on a binary 'resolved' or 'not resolved' variable that has been subjectively defined. The different outcomes of a MFH case are given in the levels of the 'return_description' variable. The 8 outcomes recorded in the data are 'arrested', 'found - deceased', 'found – family/carer', 'found – harboured and/or abducted', 'found – hospital', 'found-police', 'own accord' and 'not known/other'. Within the analysis the 'not known/other' was deemed as not resolved and all remaining outcomes were deemed resolved and combined to a binary variable, thus 'resolved' was the event of interest.

It can be seen within this variable that there are several types of resolution and it would be of benefit to understand how factors such as risk classification influence the time to resolution based on the type of outcome. This would lead to a competing risks analysis. Prentice et al summarise the problem of competing risks as the 'study of any failure process in which there is more than one distinct cause or type of failure' (1978: 541). Competing risks are often seen in clinical studies when a patient experiences the failure event, usually death, due to cause that was not the one of focus. In the MFH context, whilst not necessarily appropriate terminology, the failure or event of interest is a resolved case, the competing risks are the different 'causes' of this resolution. The causes are the several different outcomes as listed above. Of interest to an extension of this project would be to estimate the effects of risk classification on the different resolution causes.

7.3. Conclusion and Recommendations

The focus of the project was the effect of the three current MFH classifications on time taken to resolve a case. It was found that these risk classifications do in fact have a significant influence on the time to resolution, though not necessarily in the way that was expected prior to analysis. It was assumed that high risk cases would be resolved the quickest, followed in ascending order by medium risk and standard risk. Instead it was found that medium risk cases overtake this time to resolution, presumably due to the large number of cases that are classified as medium risk and consume police time and resources. Going forward, it would be advised to Lancashire Constabulary that the classification criteria when assigning a case a particular risk level, especially medium risk level, be reconsidered and a more equal distribution be given to standard and high risk where appropriate. This is to allow police resource be spread more evenly and fairly.

Due to the large number of cases, it would be recommended that more consideration be given to the way that cases relating to children and youth, in particular those under local authority care, are handled. It may be that a new procedure could be put in place between the police and the care system for responding to children that go missing from care, to prevent these cases being absorbed into the total number of all missing person cases. The large amount of these reports suggests that a different response is needed to prevent these children being repeatedly reported as missing.

In addition, this project highlighted the individual risk factors that best predict time to resolution, it would be suggested that these factors associated with longer times to resolution be higher prioritised in risk assessment in the hope that these persons be located in shorter time. Finally, the demographics and in particular the towns within Lancashire that produce the most missing person reports were identified, this information may be of benefit for targeted policing and early intervention.

References

Association of Chief Police Officers. (2005) *Guidance on the Management Recording and Investigation of Missing Persons*, National Centre for Policing Excellence.

Association of Chief Police Officers. (2013) *Interim Guidance on the Management, Recording and Investigation of Missing Persons*, College of Policing.

Association of Chief Police Officers. (2015) 'Letter to all Chief Constables Regarding Absent Definition', www.library.college.police.uk. http://library.college.police.uk/docs/appref/ACPO-Absent-letter-March-2015.pdf [04/09/2017].

Broström, G. (2012) *Event History Analysis with R,* Boca Raton, Fla., CRC Press.

Cox, D. (1972) 'Regression Models and Life-Tables', *Journal of the Royal Statistical Society,* Vol. 34, no.2, pp. 187-220.

Efron, B. (1977) The Efficiency of Cox's Likelihood Function for Censored Data, *Journal of the American Statistical Association,* Vol. 72, no.359, pp. 557-565.

Flynn, R. (2012) 'Survival Analysis', *Journal of Clinical Nursing*, Vol. 21, no.19, pp. 2789–2797.

Fox, J. & Weisburg, S. (2017) Package 'car', version 2.1-5, https://cran.r-project.org/web/packages/car/car.pdf. [05/09/2017]

Horikoshi, M. (2017) Package 'ggfortify', version 0.4.1, https://cran.r-project.org/web/packages/ggfortify/ggfortify.pdf. [05/09/2017]

Horton, N. & Kleinman, K. (2007) 'Much Ado about Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models', *The American Statistician*, Vol. 61, no.1, pp. 79–90.

Hosmer, D. & Lemeshow, S. (2000) *Applied Logistic Regression*, 2nd Edition, New York, Wiley.

Hothorn, T., Zeileis, A., Farebrother, R., Cummins, C., Millo, G. & Mitchell, D. (2017) Package 'lmtest', version 0.9-35, https://cran.r-project.org/web/packages/lmtest/lmtest.pdf. [05/09/2017]

Ihaka, R., Murrell, P., Hornik, K., Fisher, J., Stauffer, R. & Zeileis, A. (2016) Package 'colorspace', version 1.3-2, https://cran.r-project.org/web/packages/colorspace/colorspace.pdf. [05/09/2017]

Kaplan, E. & Meier, P. (1958) 'Nonparametric Estimation from Incomplete Observations', *Journal of the American Statistical Association,* Vol. 53, no.282, pp. 457-481.

Kassambara, A., Kosinski, M. & Biecek, P. (2017) Package 'survminer', version 0.4.0, https://cran.r-project.org/web/packages/survminer/survminer.pdf. [05/09/2017]

Klonowski, A. (2012) 'Report of the Independent Reviewing Officer in Relation to Child Sexual Exploitation Issues in Rochdale Metropolitan Borough Council During the Period 2006 to 2013', www.rochdale.gov.uk.

Le, C. (1997) *Applied Survival Analysis*, New York, Wiley.

Lewis-Beck, M., Bryman., A. & Futing Liao, T. (2004) 'Event History Analysis', in *The SAGE Encyclopedia of Social Science Research Methods*, Thousand Oaks, Calif., SAGE.

Madden, G., Rosalía, M., Rappoport, P. & Banerjee, A. (2016) 'A Contribution on the Nature and Treatment of Missing Data in Large Market Surveys', *Applied Economics*, Vol. 47, no.22, pp. 2179-2187.

Mills, M. (2011) *Introducing Survival and Event History Analysis*, Los Angeles, Calif., SAGE.

National Crime Agency. (2016) 'Missing Persons Data Report 2014/15', www.missingpersons.police.uk.

Newiss, G. (1999) *Missing Presumed…? The Police Response to Missing Persons*, Policing and Reducing Crime Unit, Home Office, London.

Newiss, G. (2015) 'Men More Likely to go Missing on Night out in Winter than any Other Time of Year, Kingston University Study Reveals', Kingston University, London.

Office for National Statistics. (2012) 'Ethnicity and National Identity in England and Wales: 2011', www.ons.gov.uk. https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/ethnicity/articles/ethnicityandnationalidentityinenglandandwales/2012-12-11 [05/09/2017]

Office for National Statistics. (2016) 'Overview of the UK population: February 2016', www.ons.gov.uk. https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/articles/overviewoftheukpopulation/february2016 [05/09/2017]

Perperoglou, A., Cessie, S. & Van Houwelingen, H. (2006) 'Reduced-Rank Hazard Regression for Modelling Non-Proportional Hazards', *Statistics in Medicine*, Vol. 25, no.16, pp. 2831–2845.

Prentice, R., Kalbfleisch, J., Peterson, A., Jr., Flournoy, N., Farewell, V. & Breslow, N. (1978) 'The Analysis of Failure Times in the Presence of Competing Risks', *Biometrics*, Vol. 34, no.4, pp. 541-554.

Subhash, L., Keim, J. & Solymos, P. (2017) Package 'ResourceSelection', version 0.3-2, https://cran.r-project.org/web/packages/ResourceSelection/ResourceSelection.pdf. [05/09/2017]

The APPG for Runaway and Missing Children and Adults and the APPG for Looked After Children and Care Leavers. (2012) 'Report from the Joint Inquiry into Children who go Missing from Care', www.gov.uk.

Thernau, T. & Lumley, T. (2016) Package 'survival', version 2.40-1, http://cran.irsn.fr/web/packages/survival/survival.pdf. [05/09/2017]

Therneau, T. & Grambsch, P. (2001) *Modelling Survival Data: Extending the Cox Model*, New York, Springer.

Wickham, H. & Chang, W. (2016) Package 'ggplot2', version 2.2.1, https://cran.r-project.org/web/packages/ggplot2/ggplot2.pdf. [05/09/2017]

Wickham, H. & Miller, E. (2017) Package 'haven', version 1.1.0, https://cran.r-project.org/web/packages/haven/haven.pdf. [05/09/2017]

Appendix: Table of Remaining Variables

The below table gives the name and description of all variables used in the analysis that have not been included in the existing tables.

Variable Name	Description
id_num	Identifier number to track repeat missing persons
mfh_date_created	Date that missing person record was created
mfh_date_found	Date that missing person was found
days_missing	Length of time missing in percentage of days
resolved	'Event' variable indicating if case was resolved
transferred	Case transferred to another police force
mfh_house_town	Missing person's home town
mfh_place_town	Town person is missing from
mfh_age	Continuous age of missing person