2021

An observational study using the Impact Point approach to measure the utility of digital forensic science in online child sexual exploitation cases.

Authors: Andrew R W Jackson[1] and Claire M B Gwinnett[1,*]

* Corresponding author, email: C.Gwinnett@staffs.ac.uk

Contributors: Marie Barrett[2, **], Lauren Yare[1], Hannah Vincent[1], Jasdeep Varraich[2] and Helen Poole[1]

** Contact for enquiries concerning the Home Office Impact of Forensic Science project, email: Marie.Barrett1@homeoffice.gov.uk

Affiliations: [1] Staffordshire Forensic Partnership[‡] and School of Law, Policing and Forensics, Staffordshire University.

[2] Data and Identity Directorate, Home Office.

[‡] A formal collaboration between Staffordshire Police and Staffordshire University, for details see: http://blogs.staffs.ac.uk/staffordshireforensicpartnership/

Contents

Executive Summary

Scope

This executive summary provides an overview of both the Operation Safenet study and the Impact Point (IP) approach that it was designed to trial. It also gives the key findings and recommendations drawn from that study. These findings and recommendations comprise those which are specific to the use of digital forensic methods as employed in Operation Safenet and those which concern the use of the IP approach to inform management and policy more broadly.

Overview

Staffordshire Police's Operation Safenet safeguards children via the detection of crimes of online child sexual exploitation (CSE). It is intelligence led and makes extensive use of digital forensic methods.

The Operation Safenet study, reported here, is one of six proof-of-principle studies[1]. These were initiated to test a new approach to the measurement of the impact of forensic science. That approach has identified 27 Impact Points (IPs), each with a corresponding Question Posed (QP). Each of these IPs is a location on the journey of a case through the Criminal Justice System (CJS), from crime to outcome at trial, where forensic science can make a contribution. Each QP is the question that must be answered in order to attain the IP concerned.

For each IP, three Impact Metrics (IMs) are proposed, namely *Effectiveness*, *Timeliness* and *Cost Benefit*. Respectively, these are the frequency, rapidity and economic cost benefit with which forensic science contributes to the attainment of the IP concerned. Each of the proof-of-principle studies concerns specified crime type(s), forensic science discipline(s) and IPs, and all of them focus on the investigative phase of the CJS journey.

The Operation Safenet study concerns online CSE, digital forensic methods and the seven IPs listed in Table 1. The only IMs that it is concerned with are *Effectiveness* and *Timeliness*. It uses data captured from past casefiles that concern intelligence referrals that were received by Staffordshire Police between 6 July 2016 and 16 December 2019, inclusive. Furthermore, the study made no interventions in the cases to which these files relate – making it observational and not experimental in nature. However, it has made use of variations in the captured data that occurred from case to case. Most notably, the cases from which data were captured include both those with and without early triage. In this context, early triage is triage which is commenced at the time of scene attendance, or shortly afterwards, and before any early interview would be expected to happen. (For further details, see footnote 4). In the context of Operation Safenet context, triage is any initial examination of information stored on a digital device for the purposes of the investigation of a suspected crime. It may have been carried out by a police officer. However, more commonly, it involved experts from Staffordshire Police's Digital Forensic Unit (DFU) attending the scene, or a police facility, to carry out that examination.

[1] Barrett, M (2021) *Impact of Forensic Science Project: Phase 1 Report*, Home Office

Table 1 The Impact Points (IPs) and Questions Posed (QPs) of the Operation Safenet study

Impact Point	Question Posed
Establish crime committed	Can we determine if a crime has been committed?
Identify victim	Can we determine who is the victim of this crime?
Safeguarding victims and suspects	Can we positively contribute to safeguarding victims and suspects?
Inform interview strategies	Can we determine information that will inform our interview strategy?
Admission of guilt at an earlier stage	Was evidence provided that led to a suspect admitting their guilt prior to them being charged with the offence?
Validate or refute accounts/sequence of events*	Can we validate or refute this account of events?
Referral for charging	Can evidence be provided that will directly lead to the referral of case for charging?

* This refers to the validation or refutation of the intelligence package received.

Key findings

A. Findings from the use of the Impact Point approach in the context of Operation Safenet

A.1 Forensic value

The Operation Safenet study gives proof of principal that the Impact Point approach provides detailed information about the value of forensic methods to the CJS in the area of online CSE. This is the first time that this value has been measured at this level of detail. Crucially, this approach has provided benchmark data that can be used for the purposes of policy development and operational management.

The *Effectiveness* metric provides clear evidence that digital forensic methods make highly successful contributions to the investigation of online CSE cases. The discovery of strong associations between the use of early triage and improvements in both *Effectiveness* and *Timeliness* (detailed below), demonstrates the power of the IP approach in the identification of factors that can further improve forensic value.

A.2 The effect of early triage

As described in A.3, A.4 and A.6, the choice to use early triage is strongly associated with multiple operational enhancements.

The vast majority of these enhancements can be well explained by early triage's ability to provide definitive and highly relevant information at a crucial stage of an investigation, and thereby allow better decision making. This provides a credible mechanism for the effects concerned.

However, the observational nature of the Operation Safenet study means that the possibility of a common cause effect cannot be ruled out (see 'The understanding of cause and effect' given in B.2, below, for more on this). It is also possible that the use of early triage has unwanted, and as yet undetected, side effects. There are tools that can be used in future studies to address these operationally important possibilities.

A.3 *Effectiveness* and the impact of early triage.

As shown in Table 2, when used, digital forensic methods contribute frequently to the attainment of the seven IPs under consideration, thus demonstrating their *Effectiveness*. As can be seen from that Table, there is a very noticeable improvement in *Effectiveness* that is associated with the choice to use early triage. This effect is statistically significant[2] and so is unlikely to have happened by chance alone.

Table 2 Early triage and the *Effectiveness* of digital forensic methods as found in the Operation Safenet study amongst opportunities where digital forensic methods were used.

	Percentage of opportunities where an overall contribution was made to the attainment of the Impact Point concerned	
Impact Point	**Without triage**	**With triage**
Admission of guilt at an earlier stage	22	49
Establish crime committed	84	97
Identify victim	68	93
Inform interview strategies	23	85
Referral for charging	58	84
Safeguarding victims and suspects	63	95
Validate or refute accounts/sequence of events	53	76
Range	**22 to 84**	**49 to 97**
Average (mean)	**53.0**	**82.7**

A.4 *Timeliness* and the impact of early triage.

Timeliness, as expressed by the median estimated time to answer the relevant QP, has been measured for three IPs. For each of these, amongst the cases analysed, there is a very noticeable improvement in this metric (ranging from 64.0 to 98.5 days) that is associated with the choice to use early triage. The IPs concerned are 'Admission of guilt at an earlier stage', 'Establish crime committed' and 'Inform interview strategies'. These were chosen because of their impact on the CJS and on resourcing decisions.

A.5 Place of early triage and the time between scene attendance and early interview

The typical (median) time between scene attendance and early interview was 66 minutes (22%) shorter when early triage occurred at the scene rather than elsewhere. In this context, elsewhere means, collectively, a police station or Staffordshire Police's

[2] The exact Wilcoxon signed-rank test reveals $V = 28$ and $p = 0.016$.

headquarters. Statistical testing suggests that this effect *might* not have happened by chance alone but that further work is needed to verify this finding[3].

A.6 Early triage and early interview

When early triage is used, a positive finding is associated with an increased probability of an early interview[4] and there is strong evidence that this effect is not due to chance alone[5]. One of the reasons why this is important is that admission of guilt is only possible during a formal police interview.

A.7 The time between referral and scene attendance, and the loss of digital evidence

For any given case, the receipt of the relevant intelligence report by the police is known as the referral. For each scene attendance recorded in the Operation Safenet dataset, the time between the referral and that attendance was measured. Amongst the data analysed (which was an appropriate subset of the whole), the typical (median) value of this time is shorter in instances where early triage found evidence of criminality on one or more mobile phones, tablets or SIM cards than in instances where it did not. This suggests that delays in scene attendance leads to loss of digital evidence – at least from such devices. However, the effect seen is not large enough to be statistically significant[6] given the sample size available. This indicates that chance alone is a plausible explanation for this effect.

This finding is based on cases that were referred to Staffordshire police in the years 2016 to 2019, inclusive. It therefore does not necessarily hold true outside that timeframe, as it might be affected by changes in criminality and technology that occur over time. Similarly, it does not necessarily apply to any one case as behaviour will vary from criminal to criminal. Nonetheless, this finding does have operational importance because it indicates that:

- one reason why unnecessary delay between referral and scene attendance should be avoided is that it *might* result in the loss of digital evidence and
- even if such delay occurs, that is not a reason to deny the use of digital forensic methods.

A.8 Further value in the Operation Safenet dataset

The Operation Safenet dataset is extremely rich and there is more work that could be usefully carried out to seek operationally valuable understanding from it. For example, time series plots could be made to look for trends in a number of variables of operational interest and the relative *Effectiveness* of various subdisciplines of digital forensics could be explored.

---

[3] The Mann-Whitney U test shows $W = 946$, uncorrected-$p = 0.029$, Bonferroni-corrected-p [to control familywise error rate inflation] = 0.058.

[4] For any given suspect in an Operation Safenet case, one or more formal police interviews will be conducted. The first of these is called the early interview and it takes place as soon as practicable, the majority within 24 hours after scene attendance. The purpose of an interview is to disclose to the suspect concerned what evidence has been found to date and give them an opportunity to account for the presence of that evidence. As previously stated, early triage is defined as triage that commences during or soon after scene attendance and before any early interview would be expected to happen. A positive triage finding is one that provides evidence of criminality. In this study, the only positive early triage findings that were known about were those which arose from the early triage of mobile phones, tablets or SIM cards. On occasion, early triage takes a long time to complete. When this happens, the early interview may be undertaken before the outcome of the early triage is known. Note that, for up to a maximum of two interviews per suspect, the Safenet dataset contains records of all interviews undertaken for each case that it details. Any third or subsequent interviews held with any given suspect are beyond the scope of this study.

[5] Pearson's chi-square testing reveals that this association is statistically highly significant ($\chi^2=28.63$, uncorrected-$p = 8.8 \times 10^{-8}$, Bonferroni-corrected-p [to control familywise error rate inflation] = 1.1×10^{-6}).

[6] Wilcoxon rank sum test with continuity correction reveals $W = 2433$, $p = 0.4109$.

Also, the data allows the calculation of the elapsed time between the submission of items to the DFU and the commencement of DFU work on those items. This could be an indicator of the workload of the DFU at any given time. An exploration of any association between this indicator and the frequency of early triage could provide insights into any interplay between such frequency and DFU workload.

B. Findings on the use of the Impact Point approach to inform management and policy

B.1 Successes

The Operation Safenet study has successfully used the Impact Point approach to quantify the value of forensic science in combatting online CSE and has done so in unprecedented detail. It has also demonstrated the potential of that approach in the creation of operationally beneficial knowledge via the identification of patterns in the data. Such methods could be used in future to detect trends, good practice and potential policy levers, thereby facilitating the maintenance and enhancement of performance – even in changing times.

The Operation Safenet dataset is a detailed quantitative snapshot of forensic practice in the fight against online CSE. It is well structured and, as noted in A.8, it contains more information than has thus far been analysed. It is therefore a valuable resource that can be further mined for future data-informed and evidence-led policy development.

As detailed below, lessons learnt during that dataset's creation could be used to:

- inform the automation of the data collection process;
- better understand cause and effect;
- reduce the opportunity for unconscious contextual bias;
- create performance-monitoring tools.

B.2 Lessons learnt

Data collection

The gathering of data from the Operation Safenet systems and case files was carried out manually. In doing so, it was possible to identify which systems held pertinent information about each case which enabled IPs to be addressed. The process of gathering these data has provided information about how this may be automated in the future.

The understanding of cause and effect

This study found many associations between choices that are under managerial or executive control (such as whether or not to use early triage) and outcomes of importance to the exercise of criminal justice. For each of these, there is a possibility that the association happened by chance alone. It is important to be able to detect such occurrences. This is because, if chance alone were responsible for the effect seen, any policy change driven by the observation of such an association would be expected to fail.

If the size of the effects seen and the number of data points analysed are both substantial, this tells against the plausibility of the effect seen having occurred due to chance alone. For example, this argument is applicable to the associations seen between the use of early triage and each of *Effectiveness* and *Timeliness*. Statistical methods allow the quantification of the probability of observing an effect of the size seen if it had happened by chance. These have been used to good effect in a number of instances in this study

If chance can be discounted, two possible explanations remain for any association seen between operational choices and outcomes of interest. These are that the choice made causes the effect seen or that whatever caused that choice to be chosen also caused that effect. Respectively, these explanations are called causal associations and common cause effects. The latter of these is problematic because, if common cause is responsible for the effect seen, alteration of policy to opt for one choice over the other in the hope of favouring a particular outcome will likely fail.

Well-designed experimental studies, qualitative research methods and techniques such as process analysis can each be used to provide information to aid the differentiation between causal associations and common cause effects. These can also assist in the detection of unwanted side effects of operational choices.

Unconscious contextual bias

The researchers who were responsible for deciding whether, in any particular case, a given Impact Point had been attained were also aware of the purposes of the Operation Safenet study. This introduces the possibility of unconscious contextual bias. A statistical test has been completed using a metric of investigative success that was independent of the Operation Safenet study[7]. The findings of this are consistent with those of that study. This does not prove the absence of contextual bias but it does provide a measure of reassurance that the effects of any such bias do not seem to be extreme. If implemented, recommendations 3.f and 3.g, given below, will reduce the opportunity for such bias in future work that uses the IP approach.

Performance monitoring

The three Impact Metrics – i.e., *Effectiveness*, *Timeliness* and *Cost Benefit* – are designed to measure the impact of forensic science. The Operation Safenet study is only concerned with the first two of these. However, at least for those two, it has provided evidence that:

- they are fit to measure the impact of forensic science;
- because they are rooted in the IP approach, they can do so in great and unprecedented detail;
- they can be used to detect effects of choices that are under operational control.

This means that they hold significant promise as metrics in performance monitoring. However, more work would be needed to establish how this could be done to best effect.

Furthermore, there could be benefits in harnessing the now proven merits of the Impact Point approach to create additional metrics that are specifically designed for the monitoring of performance. For example, these extra metrics could be along the lines of waiting times as used in healthcare settings.

Timeliness

Elapsed times were used to establish the *Timeliness* metric. The data that is available allows only estimates of these times to be made – limiting the precision/accuracy of this metric. This is a matter that should be addressed in future data collection.

[7] The rows of the Safenet dataset which record that digital forensic methods were used and that the investigative process reached a conclusion were considered. It was found that amongst these, investigative success is demonstrated in 67% of those rows that also record that early triage was *not* used and that this figure rises to 82% for those rows that show that early triage *was* used. Pearson's chi-square testing reveals that this association is statistically highly significant ($\chi^2 = 24.64$, $p = 6.9 \times 10^{-7}$).

Recommendations

1. **Forensic methods**
    a. Continue the use of digital forensic methods in online CSE cases.
    b. In intelligence initiated online CSE cases, undue delay between the receipt of intelligence and scene attendance should be avoided. However, if such delay occurs, the resourcing of the examination of devices for digital evidence should not be reduced.
    c. Conduct research using the IP approach to establish:
        i. how the value of forensic science can be enhanced yet further. This could be within the field of online CSE cases or beyond. For example, a pilot study could be conducted to find the forensic value of trace evidence in the detection of cases of domestic abuse or of cases of burglary committed by prolific offenders;
        ii. the utility of early triage using digital forensic methods in priority crime categories beyond online CSE, such as Rape and Serious Sexual Offences (RASSO) and Serious and Organised Crime (SOC).
    d. Use the findings from 1.c to further boost the value of forensic science.
    e. Conduct research to better understand the impact of early triage (as defined in this study) on the workload of the Digital Forensics Unit.
2. **Early triage in online CSE cases**
    a. Continue the use of early triage where it is currently used. If the timeliness of the early interview is of importance, such triage should be conducted at the scene, not elsewhere, provided that this option is available.
    b. Conduct research to:
        i. establish the cause of the operational enhancements that are associated with the choice to use early triage;
        ii. look for any unwanted side effects of that choice.
    c. Use the findings from 2.b to maximise the benefits associated with the use of early triage and roll these benefits out to future cases that, under current practice, would not benefit from them.
    d. In forces where early triage is not currently used in online CSE cases, conduct pilot studies to develop evidence-led policy on its future adoption.
3. **Future data collection**
    a. Establish a project-wide formal framework for data quality control;
    b. Ensure that the collected data are available in a file format, such as csv, that is readily accessed by the wide range of software used for data analysis (Excel, R, SPSS, Python, SAS etc).
    c. Ensure that each release of raw data has a time stamp and a unique filename. That file or raw data should be treated as immutable by the analysts. Any changes to the dataset made by them should be carried out on copies of the raw data.
    d. Consideration should be given to the project-wide adoption of a formal version control system such as git[8].
    e. Increase the number of IPs used. For example, IPs that specifically allow the measurement of victim experience and proportionality could be used to good effect.

[8] https://git-scm.com/ (accessed 29 August 2021)

f. Refinement, as needs be, of the current Questions Posed and their precise definition.
g. Create a standardised methodology for determining, in any given instance, whether any given investigative technique used, whether forensic or otherwise, has contributed to answering the Question Posed. It would seem to be prudent to do this in collaboration with the Forensic Capability Network's expert network on performance and risk, which is made up of force forensic leads.

h. Add variables to allow the:
    i. ready extension of the *Timeliness* Impact Metric across all IPs used and improve the precision/accuracy of that metric;
    ii. calculation of the *Cost Benefit* Impact Metric.
i. Minimise the opportunities for unconscious cognitive bias and, where possible, monitor and measure its impact.
j. Remove redundant variables from the dataset.
k. Widen data capture to multiple forces. This will increase the speed with which sufficient data can be collected to allow effects to be detected and aid the anonymisation of data.
l. Establish methods for the automated recording of data, with the necessary investment occurring at a national level.
m. Facilitate the addition of further variables, and/or further levels to existing variables, to better describe case characteristics and allow new research and policy development to occur.
n. Widen the list of crime types and forensic methods used.
o. Continue and extend the use of the `Tidy Data`[9] format.

4. Data analysis

a. Extend Impact Metrics to include *Cost Benefit*.
b. Make routine use of statistical methods.
c. Make routine the benchmarking against external measures of impact, such as CJS outcomes.
d. Analyse those aspects of the Operation Safenet dataset that have not yet been explored.

5. Transparency

a. Publish findings via peer reviewed journals.
b. As far as practicable, follow the tenants of reproducible research[10].
c. Consider generation of datasets with greater levels of anonymisation or which are fully synthetic[11], allowing wider distribution. This will encourage and facilitate research and development, whilst enhancing transparency and promoting reproducible research.

6. The understanding of cause and effect

a. All future work should include, by design, clear methods for distinguishing between causal associations and common cause effects, and the routine use of statistical

---

[9] Wickham, H (2014) Tidy Data, *Journal of Statistical Software*, **59**(10) pp1:23, https://doi.org/10.18637/jss.v059.i10

[10] For an introduction to reproducible research, see Alston, J M and Rick, J A (2021) A beginner's guide to conducting reproducible research, *Bulletin of the Ecological Society of America*, **102**(2) e01801 https://doi.org/10.1002/bes2.1801

[11] In this context, synthetic data are those which are fully computer generated. This generative process is informed by the properties of the observed data. Its purpose is to remove all records that could disclose identity whilst creating data that closely mimic the statistical behaviour of those observed data.

tests to help detect effects that are due to chance alone. Where possible, these methods should seek to understand the mechanism(s) that drive any causal associations found.

7. **The development and use of the IP approach to aid strategic decision making**
    a. The widespread adoption of the IP approach is recommended, as this will provide:
        i. granular evidence of the impact of forensic science;
        ii. a tool for the optimisation of forensic value via the identification of patterns, trends and best practice, the design of pilot studies and effective policy implementation.
    b. Development of experimental and observational frameworks that include both statistical and qualitative methods.
    c. Further development and funding of the research commissioning process to stimulate innovation.
8. **The development and use of the IP approach to aid operational management**
    a. The development and use of the following tools, all to be based on the IP approach:
        i. regularly updated interactive dashboards;
        ii. inter-force comparisons to find best practice;
        iii. a statistical process control methodology;
        iv. the development of new performance indicators as needs be;
        v. predictive modelling to aid operational planning.
    b. The provision of funding, as needs be, to allow the adoption of recommendation 8.a.

Conclusion

The Operation Safenet study has demonstrated the utility of the Impact Point approach in the measurement of forensic value. It has identified a number of choices that are under operational control that are associated with the enhanced performance of digital forensic methods in online CSE cases. Most notable amongst these choices is the decision, in any given case, whether to deploy early triage. Using the Impact Point approach, it is clear there are significant benefits associated with the use of early triage for CSE cases[12]. It would seem likely that early triage using digital forensic methods could be of merit in crime categories such as RASSO and SOC, and research to explore this potential would be of value. The Operation Safenet study has also demonstrated the potential of the Impact Point approach as a machine for knowledge creation to inform both policy development and management purposes. The recommendations given above provide a route to the realisation of that potential.

[12] Although, as discussed in B.2, it has not been proven that these benefits have been caused by the use of early triage.

Acknowledgements

This work is funded by the Home Office, and without this support it would not have been possible. Thanks are due to Staffordshire Police in general, and John Beckwith, Claire Millar and Chris Hood in particular, for assistance in the collection of the Operation Safenet data. Thanks are also due to Carolyn Lovell who was a member of the project guidance team and helped in the design and implementation of the study.

List of abbreviations

ACPO	Association of Chief Police Officers
CJS	criminal justice system
CSE	child sexual exploitation
DFU	Digital Forensics Unit
FAO	first officer attending
IM	Impact Metric
IP	Impact Point
OIC	officer in charge
QP	Question Posed
SOCO	Scene of Crime Officer

1. Introduction

1.1 Effectiveness of Forensic Science in Criminal Investigations

Forensic science is vital to the investigation of crime and the efficient and effective operation of the criminal justice system (CJS) because of the support it provides to both the police and courts (Woodman et al., 2019). This is particularly true for many serious and major crimes, including those that are classed as high harm offences such as child sexual exploitation (CSE). The processes used to detect, obtain and present forensic evidence are complex, and vary dependent on many factors including but not limited to, case type, evidence found, and police and forensic provider standard operating procedures (Wüllenwebe and Giles, 2021).

Although those working in the criminal justice system anecdotally know the benefits of using forensic science in investigating crimes, there is still minimal data on its true impact on criminal investigations. A review of literature, from the 1980s to 2014 by Ludwig and Fraser (2014), found limited formal research that has been undertaken which provides empirical data on the effectiveness of forensic science in criminal investigations. Due to this lack of research, there is still a need to provide a robust evidence base of the impact that utilising forensic approaches can have on a case.

It has been contended that better understanding of how forensic science contributes to criminal investigations can lead to improvements in its strategic use. This is particularly important when forensic resources are limited and demand for services is high. It has been proposed that improved forensic strategies also lead to other benefits, including reductions in backlogs, higher clearance rates of offences and, ultimately, reductions in crime rates (Ludwig and Fraser, 2014).

Evaluating how effectively forensic science contributes to criminal investigations is complex. Woodman et al. (2019) explain that part of that complexity is due to the variation seen across the spectrum of forensic disciplines, including whether the services are conducted at the scene, in the laboratory or both, and differences in the typical time taken to complete the work from discipline to discipline. Furthermore, forensic science will impact on the criminal justice system in different ways dependent on the type of science involved. Therefore, specific studies are required for each scientific discipline, whilst acknowledging that criminal cases rarely utilise solely one type of evidence or approach (McLaughlin, 2021).

It has been noted that, although there is a significant amount of literature on the effectiveness of forensic science, the number of research studies with robust experimental design is limited (Ludwig and Fraser, 2014). Examples of these research studies include Roman et al.'s (2015) field experiments on the use of DNA in high volume crimes and Wüllenweber and Giles's (2021) assessment of the contribution of forensic evidence to the acquisition of criminal charges in volume crimes. Ludwig and Fraser (2014) go on to highlight some of the limitations of the current literature, including a lack of statistical analysis and evaluation, and the lack of rigour in the methodological approaches used. A recent review conducted by McLaughlin (2021) as part of the wider project that this study belongs to (Barrett 2021a), also produced a comprehensive overview of the past literature and highlighted that most research has been on factors impacting *on* forensic science rather than the impact *of* forensic science. The first attempts to understand the effectiveness and efficiency of forensic science include the Touche Ross report in the UK (Touche Ross, 1987) and the Peterson et al. report (1984) in the US. Relevant published research focussing on the impact of specific forensic disciplines on case investigations includes work on: the contribution of chemical traces to criminal justice outcomes (Woodman et al., 2019), the use of CCTV in murder investigations for offender identification (Brookman et al., 2020), the use of DNA in solving property crimes (Roman et al, 2009) and the impact of forensic anthropology techniques on the identification of individuals (Uberlaker, Shamlou and Kunkle, 2018).

Studies that have examined the impact of multiple forensic science disciplines on the investigative process include: a US study of such impact on case-processing decisions (Peterson et al., 2013) and UK studies on the impact on volume crime (Ludwig and Fraser, 2014; Wüllenwebe and Giles, 2021). When measuring effectiveness, these studies have utilised a number of indicators of performance, including, but not limited to, timeliness of results, case turn-around times and the efficiency of case progression to court.

We contend that the measurement of the effectiveness of forensic science should be ongoing, incorporate multiple evidence types and be based on a significant number and range of cases. This is in the belief that, ultimately, if appropriately employed, such measurement would allow evidence-based development of operational police policy, and better-informed budgeting and resourcing decisions. Also, it would allow the measurement of the impact of those decisions and of policy implementation. To date, there has been no study published that describes the design and execution of a method of assessing effectiveness in this manner and to this effect.

The study reported here focusses on the effectiveness of digital forensics and of the use of early triage in the investigation of online child sexual exploitation (CSE) cases. Until now, there has been no attempt at quantifying effectiveness of digital forensics in this context.

1.2 The increased need for digital forensic services in child sexual exploitation (CSE) cases

In the United Kingdom, child sexual exploitation (CSE) has been defined by the Home Office (2016a) as a form of sexual exploitation where a child under the age of 18, is sexually abused for power, status or money. Under section 160 of the Criminal Justice Act 2003, the possession of an indecent image of a child (IIOC) is a criminal act, as is the taking, distribution or sharing of one or more IIOCs (Kloess, Woodhams and Hamilton-Giachritsis, 2021). The sexual exploitation of children has occurred for almost certainly throughout all of human existence. However, since technology advancements have led to the prolific use of smartphones, computers, mobile applications and social media, there has been a rise in CSE cases (Kloess et al., 2019). The world wide web has allowed for the global distribution of and easy access to IIOCs. Owens *et al.* (2016), state that the internet also allows for anonymity where offenders are able to access the material in their own home with relativity low risk of being detected. Between 1996 and 2007, there was a 1404% increase in CSE cases in the UK and it has been estimated that there were 1 million IIOCs circulating in 2003 which has increased by an additional 200 photographs every day since then (Jung, *et al.* 2013). Consequently, this influences the amount of evidence that is submitted for digital forensic examination and therefore contributes to backlogs within digital forensic units (Ngejene, *et al.* 2021).

Brookman and Jones (2019) investigated the role of forensic science in murder investigations. They noted that digital forensic methods, such as mobile phone analysis, were used more regularly than other forms of forensic evidence gathering, including DNA analysis. It was proposed that this was due to the results being available more quickly from these digital methods, thus leading to detectives giving them a higher emphasis. It would seem reasonable to conclude that this experience of 'fast results', and the ever-increasing use of digital devices and media generally, is likely to also put pressure on Digital Forensic Units.

1.3 The use of triage in digital forensics

Digital triage can be described broadly as a process which helps ensure that all resources that are available to the investigative team are utilised efficiently. It has been noted by Shaw and Browne (2013) that the definition of triage is generally poorly explained and consequently has different meanings to multiple individuals. It is generally understood though that the aim of triage is to rapidly locate evidence, individuals who may be at risk and suspects. Mislan, Casey and Kessler(2010) describe the process of digital triage as having the following steps; locating any devices that may be evidentially useful, conducting a preliminary examination of those items to provide information quickly to the investigators to aid with the development of leads and to be used in interviews, and finally the recovery of prioritized items for further, more in-depth analysis.

Disagreements as to whether triage should be conducted in the field (at scene), referred to as 'pre-seizure'[13], or at the laboratory, referred to as 'post seizure', exist. Rogers et al. (2006) state triage should be field based only, whilst Martell, Quates and Roussev (2013) note it can be conducted in either location – as long as it is carried out by a digital expert. CSE cases are deemed time critical, therefore deciding where triage should be conducted is a vital component for the investigation. Concerns have been raised that pre-seizure triage might result in evidence being lost (Horsman, Laing and Vickers, 2018; Mislan, Casey and Kessler, 2010). The reasoning being that equipment available for scene use might be of lower quality than that in the laboratory and evidentially valuable items might be wrongly deemed to be too low a priority to seize. Given the time sensitive nature of CSE cases, the investigative team need to balance these concerns against the delay caused by transportation when triage is conducted post seizure. This time delay may impact on whether any initial evidence can be presented to a suspect at an early interview. This ultimately may affect the outcome of that interview as Moston and Engelberg (2011) state there is a higher probability that a suspect will admit guilt in the presence of evidence. In addition, Mislan, Casey and Kessler (2010) note that evidence may be lost during transportation due to power loss or remote wiping.

There is a general agreement as to the maximum time for conducting 'at scene triage', which should ideally be completed within 60 minutes but no more than a few hours from the time of arrival at the scene (Martell, Quates and Roussev, 2013; Shiaeles, Chryssanthou, and Katos 2013; Nickson and Venter, 2015). Nickson and Venter (2015) propose that conducting triage in under an hour can reduce the stress experienced by the suspect whilst their devices are being searched, and could provide a better duty of care and safeguarding to the suspect especially when dealing with young offenders. They indicated that conducting triage in this timeframe could potentially have a less negative effect on the suspect, making them more cooperative when it comes to further investigations. Conversely, Horsman, Laing and Vickers (2014) suggest that using a short time-frame to examine the evidence could be a disadvantage, especially if the scene has multiple devices. Their reasoning being that time pressure might mean that devices would not undergo as rigorous an examination, which could lead to evidence being missed, or result in more items having to be seized – thereby adding to any backlog of work in the DFU.

At-scene triage may involve the use of specialist technology such as FDumper, TriageIR (Bashir and Khan, 2015) and SEAKER (Gentry and Soltys, 2019) or be conducted through the application of strategic decision making led by the first officer attending (FAO), also known as decision-based device triage (DBDT). The latter involves evaluating the evidential value of each device and deciding whether to seize it based on a series of assumed or known facts about the item such as its ownership and perceived usage (Horsman. 2020).

[13] In the study reported here, the term used for pre-seizure triage is early triage.

There are many potential benefits of using early triage in digital investigations of CSE cases, these include:

- **Reduction in the amount of digital evidence to be analysed.** Triage has been described as a solution to the large amount of data that needs to be analysed in a case (Shaw and Browne, 2013) due to its ability to prioritize key evidence for further analysis.
- **Rapid elimination of devices and individuals believed to be involved in a crime.** Triage is designed to aid in creating an efficient criminal investigation, which the Association of Chief Police Officers (ACPO) and the Forensic Science Service once described as not only leading to the identification and prosecution of offenders but also the rapid elimination of innocent individuals (ACPO and FSS, 1996). As digital triage allows items to be analysed at the scene, those which do not have a positive result for the presence of evidence related to a CSE case can be returned to their owners and individuals may be quickly eliminated as suspects.
- **Improved communication between digital experts and police.** Horvath and Meesig (1996) describe that police require feedback from forensic scientists on the suitability of evidence for certain outcomes. When digital experts attend the scene, information about appropriate sources of evidence and their value can easily be communicated. This knowledge transfer concerning forensic science value has been noted as being rare (Ludwig and Fraser, 2014).
- **Improved searching and recovery of evidence.** Multiple reports have described the issues of limited knowledge of forensic science by police officers (Blakey, 2000; Horvat and Meesig, 1996; Ludwig and Fraser, 2014). When sending police officers to a CSE scene, they may not be well equipped to identify sources of digital evidence, especially items which are have purposely been disguised or are unusual in appearance. Conversely, digital experts are experienced in a multitude of device types and will be able to interpret what they find within the context of the investigation at the scene.

In a paper discussing forensic science provision, Roux et al. (2015) highlight the need for a more holistic approach to investigating crimes. It could be argued that the use of early triage by sending DFU experts to the scene would be a step towards creating this more holistic approach to meeting the needs of the criminal justice system and maximising the utility of digital evidence.

1.4 Operation Safenet

Operation Safenet is a dedicated and proactive online child sexual exploitation (CSE) investigation process, which was created in July 2015 by Staffordshire Police. That operation utilises expert digital forensic interventions at the scene of warrants, which are executed following the receipt of intelligence of online CSE. It aims to utilise the latest technology to gather evidence and speed up the police response in order to protect children at risk. (Cooper 2021)

1.5 Project overview and aims

The need to better understand forensic value has been recognised for some time. The Home Office's 2016 Forensic Science Strategy states:

"There is a need for in-depth analyses to enhance our understanding of the specific contribution of forensic science to the CJS [the Criminal Justice System] in England and Wales" (Home Office 2016b)

In a similar vein, the 2018 (published 2019) joint forensic provision review asserts:

"While assessing the impact of forensic evidence is challenging, some measures to indicate its value to criminal justice outcomes would be strongly preferable to reliance on anecdotal feedback." (Home Office, APCC and NPCC 2019)

In response to this need, in 2020, the Home Office initiated the Impact of Forensic Science project to develop a methodology for the quantitative measurement of the value of forensic science across the CJS (Barrett, 2021a). Its approach is to identify Impact Points (IPs), each with a corresponding Question Posed (QP). Each IP is a location in the journey of a case through the CJS where forensic science can make a contribution. Each QP is the question that must be answered in order to attain the IP concerned. The concept is then to gauge, for any given case, whether the forensic and non-forensic methods used contributed to the answering of each QP and, if so, the time and cost that it took to do this. By these means, three Impact Metrics (IMs) – namely *Effectiveness*, *Timeliness* and *Cost Benefit* – would be created for each IP. Respectively, these are the frequency, rapidity and economic cost benefit with which forensic science contributes to the attainment of the IP concerned. The complete set of IPs and QPs that have been identified by the Impact of Forensic Science project is given in Appendix 1 (supplied by Barrett, 2021b).

The Impact of Forensic Science project also contains six proof of principle studies (Barrett, 2021a). The purpose of these is to test whether the Impact Point approach, described above, offers a clear and useful set of metrics for the measurement of forensic value. If it proves to be a successful approach, that project also will begin to make the case its use based on manually collected data to inform strategic and operation decision making. More importantly, it has started to argue for the development of the current CJS crime and case management systems to allow real-time automated generation of metrics of forensic value. Access to this type of information will allow the development of a data- and evidence-driven model of operational decision making and of policy development and implementation. It would also allow the identification of operational best practice in forensic science delivery and pertinent trends.

One of the afore-mentioned proof of principle studies is the Safenet study (also known as the Operation Safenet study), which is the subject of this report. It utilises the Staffordshire Forensic Partnership, which is a formal collaboration between Staffordshire Police and Staffordshire University[14]. It centres on the use of data drawn from Operation Safenet cases. Its primary purpose is to assess the utility of the Impact Point approach via the application of the *Effectiveness* and *Timeliness* impact metrics using the Impact Points shown in Table 1.1. Due to the nature of Operation Safenet (described in Section 1.4), the Safenet study is concerned with cases of online CSE and the use of digital forensic methods.

[14] http://blogs.staffs.ac.uk/staffordshireforensicpartnership/ (accessed 24 August 2021)

Table 1.1 The Impact Points (IPs) and Questions Posed (QPs) of the Safenet study (See Appendix 1 [supplied by Barrett (2021b)] for a list of all 27 of the IPs and their corresponding QPs as used in the Impact of Forensic Science project as a whole)

Impact Point	Question Posed
Establish crime committed†,‡	Can we determine if a crime has been committed?
Identify victim†	Can we determine who is the victim of this crime?
Safeguarding victims and suspects†	Can we positively contribute to safeguarding victims and suspects?
Inform interview strategies†,‡	Can we determine information that will inform our interview strategy?
Admission of guilt at an earlier stage†,‡	Was evidence provided that led to a suspect admitting their guilt prior to them being charged with the offence?
Validate or refute accounts/sequence of events†,*	Can we validate or refute this account of events?
Referral for charging†	Can evidence be provided that will directly lead to the referral of case for charging?

* This refers to the validation or refutation of the intelligence package received.
† Used to measure *Effectiveness*
‡ Used to measure *Timeliness*

The Safenet study also has a secondary purpose. This is to make use of variation in the characteristics of the cases that it has sampled to search for the existence of specific associations of operational importance that might be present in the data. In particular, advantage was taken of the fact that early triage[15] was not always used during the processing of the cases of interest. For example, this has allowed the possibility of the association between the use of such triage and *Effectiveness* and *Timeliness* across a range of IPs to be explored.

Use of variation in case characteristics has also permitted the following specific research questions to be addressed:

- Does place of early triage have an effect on the time between scene attendance and early interview? (See Sections 2.2.2 and 3.2)

[15] Here, the term triage means the initial examination of information held on digital devices found at the scene of warrant execution. By definition, early triage is that which is commenced at the time of scene attendance, or shortly afterwards, and before any early interview would be expected to happen. Early triage may be undertaken by digital forensic specialist staff or police officers. It is noteworthy, however, that amongst the 1398 observations (i.e., rows) of the Safenet dataset in which it is recorded that early triage occurred, 1329 (95%) also record the deployment of such specialist staff for the purposes of carrying out such triage. For more detailed information on the meaning of early triage, see the definition of the 'At Scene triage' case characteristic that is provided in Appendix 2. For more information on the nature of the early interview, see footnote 4.

- Does early triage lead to early admission of guilt? (See Sections 2.2.3 and 3.3)
- Does delay between the date of referral and the date of scene attendance lead to loss of digital evidence? (See Sections 2.2.4 and 3.4)

Work such as this has not only provided insights of operational value but has also allowed the potential of the Impact Point approach to be used to facilitate evidence-based and data-driven policy decisions to be demonstrated (see Section 4).

2 Methods

Ethics was approved via Staffordshire University's Research Ethics board.

2.1 Data collection

The Safenet dataset is a spreadsheet made up of data gathered from police records for all 243 Operation Safenet cases that were notified to Staffordshire Police in the period 6 July 2016 to 16 December 2019. Each of these cases concerns suspected online child sexual exploitation within the geographical area covered by that police service.

For every case, those records were used to gather the case characteristics listed in Table 2.1. Many of these characteristics are defined in Appendix 2. Drop-down menus were used to provide specified values for some of the case characteristics; these are provided in Appendix 3.

Table 2.1 Case characteristics and information categories

Information category	**Case characteristic**
Pre scene attendance	Date referral received
	Source of referral
	KIRAT Level
	Overall Risk Assessment
Scene Attendance	Police scene attendance
	Date of scene attendance
	Time of scene attendance
	Have DFU staff attended?
	How many DFU staff attended?
	How many mobile devices were seized for initial triage examination?
	How many tablets were seized for initial triage examination?
	How many Computers were seized for initial triage examination?
	How many other exhibits were seized for initial triage examination?
	How many mobile devices were examined at the scene?
	How many tablets were examined at the scene?
	How many computers were examined at the scene?
	How many other digital devices were examined at the scene?
	How many mobile devices were recovered from the scene?
	How many tablets were recovered from the scene?
	How many computers were recovered from the scene?
	How many other digital devices were recovered from the scene?
	Early evidence provided
	Referral package intelligence confirmed
	First generation images identified

Information category	Case characteristic
Early investigation	Early interview taken place?
	Date of early interview
	Time of early interview
	Place of interview
	Early interview suspect account
	Early investigation outcome
	Support referral
DFU Laboratory	Date of DFU submission – MG21 to submissions
	How many mobile devices submitted for DFU examination?
	How many tablets submitted for DFU examination?
	How many computers submitted for DFU examination?
	How many other exhibits submitted for DFU examination?
	Date DFU work started
	Date DFU work completed
	Number of mobiles devices examined in DFU
	Number of tablets examined in DFU
	Number of computers examined in DFU
	Number of other exhibits examined in DFU
	Number of mobile devices eliminated by laboratory triage
	Number of tablets eliminated by laboratory triage
	Number of computers eliminated by laboratory triage
	Number of other exhibits eliminated by laboratory triage
	Referral package intelligence confirmed
	SFR produced
Investigation	Further interviews taken place?
	Date of interview
	Time of interview
	Place of interview
	Interview suspect account
	Interview outcome
	Support referral
	CPS referral made
Disposal	Charges applied
	Charges
	Early case management SFR agreed?
	Guilty plea
	Trial court
	Trial outcome
	Final Sanction
Approach taken to answer the question	Was forensics used?
	Forensic discipline 1
	Did it contribute
	Forensic discipline 2
	Did it contribute
	Forensic discipline 3
	Did it contribute
	Forensic discipline 4
	Did it contribute

Information category	Case characteristic
Approach taken to answer the question	Forensic discipline 5
	Did it contribute
	Forensic discipline 6
	Did it contribute
	Did forensic contribute overall?
	Did forensics contribute exclusively?
	Non-forensic approach used
	Type of non-forensic approach used
	Type of non-forensic approach used
	Type of non-forensic approach used
	Type of non-forensic approach used
	Did any non-forensic approach contribute?
	Did the non-forensic approach contribute exclusively?
	Did both methods contribute?
	Estimated time taken to answer the question
	Estimated cost of answering the question
None	Final Outcome
	At scene triage
	Forensic not started
	Place of triage

The case characteristics, listed in Table 2.1, form columns of the Safenet dataset. To obtain the information needed to populate those columns, the case management systems listed in Figure 2.1 were used. As indicated in that Figure, each of these systems was accessed in the order shown and contained specific information about each case.

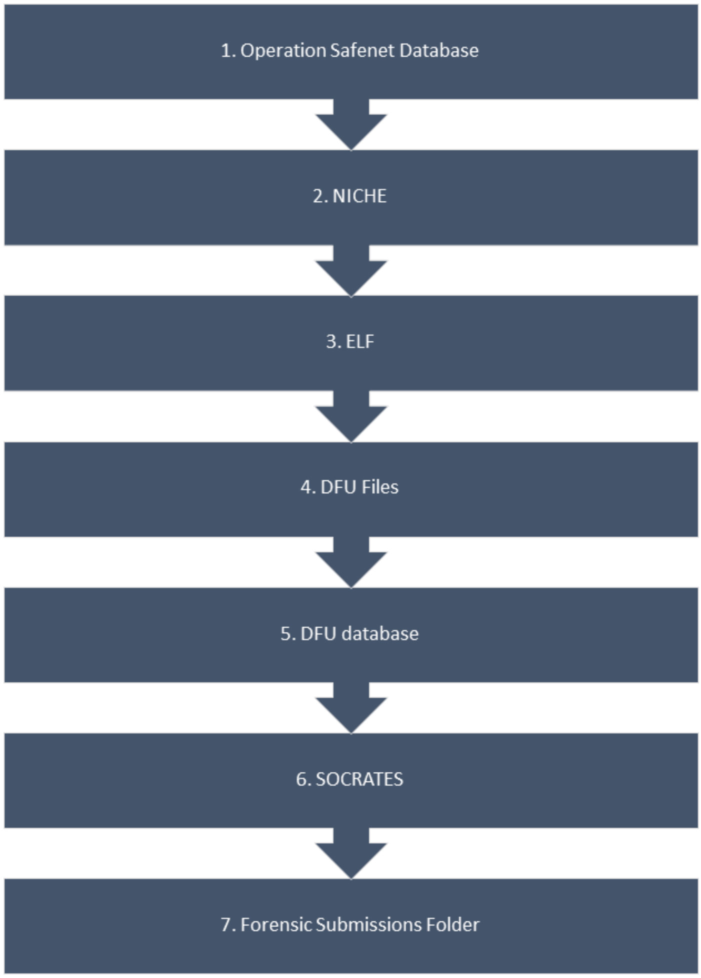


Figure 2.1: Order of access to systems used during data collection.

Details of which of the systems named in Figure 2.1 were used to identify each of the case characteristics listed in Table 2.1 can be found in Appendix 4.

Due to it being an active database, no direct access to the Operation Safenet Database was possible. Instead, a copy of the data was provided where only cases that were noted as 'actioned' and from between 2016 and 2019 were selected. In this context, 'actioned' means that police enforcement at the suspect's location had taken place.

Data was gathered from NicheRMS by searching by the name, date of birth and occasionally the address of the suspect in the case. Upon confirming that the correct case was being accessed, the case details would be found in the occurrence enquiry log.

Data was gathered from ELF by entering the collar number of the Officer in the Case (OIC) and then searching for the surname of the suspect.

The physical DFU files were used for gathering information on early triage for mobile phone evidence only. Cases were identified via the surname of the suspect and date actioned on the case file cover.

The DFU database was only utilised when there was limited information on the other systems due to access issues due to the database being live.

Data was gathered from SOCRATES by selecting Operation Safenet operations only and then selecting a specific case by 'date actioned' and confirming the case via the suspect's name and address.

Finally, the Forensic Submissions Folder was used to access the MG12 form for each case by entering the Scene of Crime Officer (SOCO) file number for a particular case.

Certain characteristics required the confirmation of a DFU officer, these are noted in Appendix 4.

For each case in the Safenet dataset, there is one row for each opportunity that digital forensics had to attain one of the seven Impact Points (IPs) by addressing the corresponding Question Posed (QP) (for details of the IPs and QPs, see Section 1.5 and Appendix 1). All cases offer one or more opportunities for this to happen, meaning that all cases have at least seven rows in that dataset (Figure 2.2). Appendix 4 shows the source of information used to determine whether any given QP has been addressed.

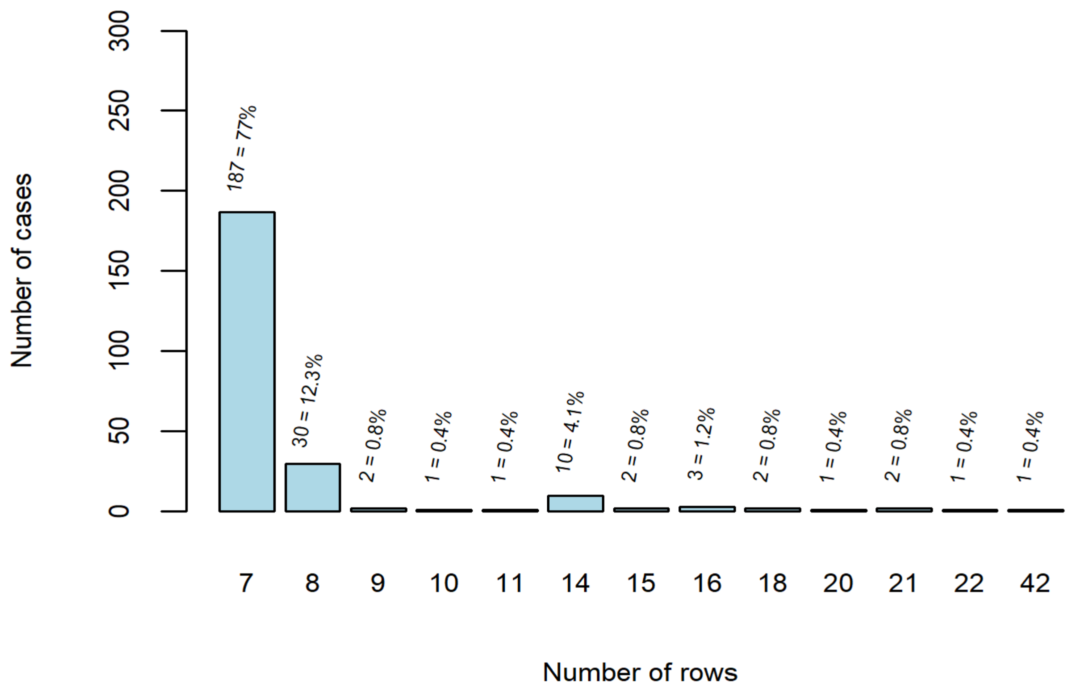


Figure 2.2 Bar chart showing the frequency of cases when categorised by the number of rows in which any given case appears in the Safenet dataset

In addition to the case characteristics, columns of the Safenet dataset were used to record information such as the relevant IP and QP for each of its rows. In total, the Safenet dataset is a spreadsheet of 1968 rows (observations) and 96 columns (variables). The names of those variables are given in Appendix 5, the values of 74 of which are summarised in Appendix 6.

The retrospective nature of the Safenet data has a number of implications, notably:

- It was possible to gather data from cases that were initiated over a time period that is much longer than the duration of the study. This has allowed the collection of a substantial amount of data.
- the inevitable disruptive effects of the COVID-19 pandemic have not perturbed the data in any way;
- there can be no contextual bias caused by the Safenet study in the *raw* data. This is because all of those data were created prior to its existence. (See Section 3.1.3 for a discussion of this topic and how, despite this fact, the possibility of contextual bias remains).
- the Safenet study is observational and not experimental in nature. (See Section 4.1.2 for more on this and the implications that this has for the understanding of cause and effect).

After the completion of the data gathering phase, the Safenet dataset was copied and that copy was pre-processed ahead of the analysis that is described in Section 2.2.1. That pre-processing detected a few data entry errors which were corrected by hand, both in the copy and in the final version of the Safenet dataset. That corrected final version is the one dated 14 May 2021 that was used for the sub-studies described in Sections 2.2.2, 2.2.3 and 2.2.4

2.2 Data analysis

2.2.1 Dashboards from the Safenet dataset

Effectiveness and *Timeliness* are Impact Metrics of the previously mentioned Home Office project (Barrett 2021a, 2021b). For any given Impact Point (IP), these are, respectably, the frequency and rapidity with which forensic science contributes to the attainment of the IP concerned. In the Safenet study, Barrett and Varraich (2021) used Excel to:

- measure *Effectiveness* for each of the seven Impact Points (IPs) listed in Table 1.1;
- estimate *Timeliness* for three of those IPs (namely, Admission of guilt and an earlier stage', 'Establish crime committed' and 'Inform interview strategies').

As indicated above, the forensic science used in the Safenet study is that provided by digital forensic methods.

Amongst variables captured in the columns of the Safenet dataset there are those which, for each row, identify:

- which case the row concerns (this is done via anonymised case reference numbers);
- whether each of forensic and non-forensic methods were used. In the context of the Safenet study, any given method (whether forensic or not) is deemed to have been used only if the method concerned both:
  a) has the potential to be able to contribute to the answering of the relevant Question Posed and
  b) was actually deployed to address that question during the investigation of the case concerned;
- whether triage was used. Here, triage is what is also referred to as early triage elsewhere in this report. As previously defined, it is triage which occurred at the time of scene attendance, or shortly afterwards, and before any early interview would be expected to happen. For further details, please see footnotes 4 and 16;
- the IP and QP of the row in question;
- whether digital forensic methods made a contribution to the answering of the relevant QP and, if so, which forensic disciplines were involved and whether those methods did so exclusively or in combination with non-forensic methods;
- whether first generation images were identified;

- the KIRAT level. This is the perceived level of risk of the suspect carrying out contact offences against children as assigned using the Kent Internet Risk Assessment Tool. This tool recognises four categories of this risk, namely: very high, high, medium and low. For more on KIRAT, see Long et al. (2016);
- whether Digital Forensics Unit (DFU) staff attended the scene, a police station or the police HQ for the specific purpose of carrying out early triage;
- the final outcome of the investigation;
- key dates and times during the investigation. Data of this type were used by Barrett and Varraich (2021) to estimate *Timeliness* for the IPs 'Admission of guilt and an earlier stage', 'Establish crime committed' and 'Inform interview strategies'. They did this by using the rules that they devised which are set out in Appendix 7.

By employing Excel's pivot table and pivot chart functions, this information was used by Barrett and Varraich (2021) to create five synoptic dashboards[16] from the Safenet dataset. These are named 'Overview', 'Contribution by Impact Point', 'Triage comparison', '*Timeliness*' and '*Timeliness* (Triage Comparison)' and their key features are summarised in Tables 2.2, 2.3 and 2.4. The first three on that list of dashboards display *Effectiveness,* whilst, as would be expected from their names, the others show *Timeliness*. Collectively, they allow the user to see how these two metrics vary dependent on whether early triage was used (see Section 3.1 for details).

[16] Most dashboard applications are updated with new data on a regular basis. There are currently no plans to carry out such updating for the dashboards described in this report because the data-collection process has now ceased.

Table 2.2 Data *excluded* prior to dashboard creation.

<table>
<tr><th colspan="5">Dashboard*</th></tr>
<tr><th>Overview</th><th>Contribution by Impact Point</th><th>Triage comparison</th><th>Timeliness</th><th>Timeliness (Triage Comparison)</th></tr>
<tr><td colspan="5">For all dashboards, rows with both no forensic methods used and no non-forensic methods used (this includes rows where the relevant variables have been left blank)

The row for one case that contained the ‘Establish Crime Committed’ IP. This was excluded because:
• the digital evidence that was seized was not examined and was sent to another force, and
• no non-forensic methods were using in this instance.</td></tr>
<tr><td></td><td></td><td>Cases where forensic methods were not used.</td><td colspan="2">Cases where forensic methods were not used.

Observations for the following Impact Points: iv, svs, vras and rc[†].

For two cases, rows that contain the ‘Inform Interview Strategy’ IP. This is because, in these instances, the date on which the DFU work was completed is unknown[‡].

For one case, the row that contains the ‘Establish Crime Committed’ IP. This is because, in this instance, the date of submission of item(s) to the DFU and the date on which the DFU work was completed are unknown[‡].

All rows that contain the ‘Admission of Guilt at an Earlier Stage’ IP and where suspect has not admitted allegations in either interview[‡, **].</td></tr>
</table>

* In this Table, as elsewhere in this report, IP stands for Impact Point (see Section 1.5) and DFU stands for Staffordshire Police’s Digital Forensics Unit.
[†] iv = Identify victim, svs = Safeguarding victims and suspects, vras = Validate or refute accounts/sequence of events and rc = Referral for charging.
[‡]As per the *Timeliness* rules that were devised by Barrett and Varraich (2021) and which are set out in Appendix 7.
**For up to a maximum of two interviews per suspect, the Safenet dataset contains records of all interviews undertaken for each case that it details. Any third or subsequent interviews held with any given suspect are beyond the scope of this study.

Table 2.3 Data displayed in each dashboard

Type of data	Dashboard				
	Overview	**Contribution by Impact Point**	**Triage comparison**	***Timeliness***	***Timeliness* (Triage Comparison)**
Number of cases	Data displayed (239 in total)	Data displayed (239 in total)	Not displayed	Data displayed (215 in total)	Data displayed (215 in total)
Number of Impact Points (i.e., observations)	Data displayed (1258 in total)	Data displayed (1258 in total)	Data displayed (1129 in total)	Data displayed (497 in total)	Data displayed (497 in total)
Date range	Data displayed (06/07/16 to 16/12/19 in full)	Data displayed (06/07/16 to 16/12/19 in full)	Not displayed	Data displayed (06/07/16 to 13/12/19 in full)	Data displayed (06/07/16 to 16/12/19 in full)
Did forensics contribute overall? (count and percentage)	Data displayed (Sum totals. Table and pie chart)	Data displayed (Sum totals and grouped by IP. Table and bar chart)	Data displayed (Grouped by triage or no triage and IP. Table and bar chart)	Not displayed	Not displayed
Did forensics contribute exclusively? (count and percentage)	Data displayed (Table and pie chart)	Data displayed (Sum totals and grouped by IP. Table and bar chart)	Not displayed	Not displayed	Not displayed
Did forensics contribute – grouped by forensic discipline? (count and percentage)	Data displayed (Table and bar chart)	Not displayed	Not displayed	Not displayed	Not displayed
Count and percentage of each category of final outcome	Data displayed (Table and pie chart)	Not displayed	Not displayed	Not displayed	Not displayed
Estimated time taken to answer the Question Posed[1]	Not displayed	Not displayed	Not displayed	Data displayed (Overall and grouped by IP[†]. Descriptive statistics, and box and whisker plots[‡])	Data displayed (Overall and grouped by whether triage occurred. Descriptive statistics, and box and whisker plots[‡])

[†]Only the given for the following IPs: 'Admission of guilt at an earlier stage', 'Establish crime committed' and 'Inform interview strategies'.
[‡]For information on box and whisker plots, and their interpretation, see Galarnyk (2018)

Table 2.4 Filters available for each dashboard

Filter	Dashboard				
	Overview	**Contribution by Impact Point**	**Triage comparison**	***Timeliness***	***Timeliness* (Triage Comparison)**
Was forensics used?	Available	Available	Not available	Not available	Not available
Did forensics contribute overall?	Available	Available	Not available	Available	Available
Did forensics contribute exclusively?	Available	Available	Not available	Available	Available
Impact Point	Available	Not available	Not available	Not available	Available
All four combinations of triage use (Yes/No) and forensics use (Yes/No)	Available	Available	Not available	Available	Not available
Final outcome	Available	Available	Not available	Available	Available
First generation images identified	Not available	Not available	Not available	Available	Available
KIRAT level	Not available	Not available	Not available	Available	Available
Have DFU staff attended?*	Not available	Not available	Not available	Available	Not available

*This refers to the attendance of staff from the Digital Forensics Unit at one of the scene of warrant execution, a police station or the police HQ for the specific purpose of carrying out early triage for the case concerned.

2.2.2 Place of early triage and the time between scene attendance and early interview

This sub-study was conducted to address the research question: 'Does place of early triage have an effect on the time between scene attendance and early interview?'. It used a copy of the version of the Safenet dataset that is dated 14 May 2021. Prior to analysis, that copy was processed to retain only one row per unique combination of case reference number, and date and time of scene attendance. This was done to avoid artificially inflating the sample size (i.e., the number of rows) and to ensure that neither the under- nor the over-representation of cases occurred. To do this, for each case reference number, those rows with duplicate values of the date and time of scene attendance were removed.

Then, of the remaining rows, all were excluded that contained records showing one or more of the following:

- that early triage did not occur;
- no record of where any early triage happened;
- no record of the date and time of either the scene attendance and/or an early interview.

For the rows that were left, the time, T, between scene attendance and early interview was calculated. Any rows in which it is recorded that the early interview occurred before the scene attendance would have then been removed. However, this was not necessary as there were no such instances found.

The processing of the data as detailed above reduced the number of cases in the copy from 243 to 105.

Each instance of early triage that is recorded in the dataset happened at one of three possible places, namely the scene, a police station or at Staffordshire Police's headquarters. Two versions of the processed data were created. In one, these locations were labelled as "Scene", "Station" and "HQ", respectively. In the other, two location labels were used, one, as before, was "Scene", with the other being "Elsewhere". The latter of which was used to indicate that early triage happened at either a police station or Staffordshire Police's headquarters.

Descriptive statistics have been calculated and box and whisker plots have been drawn[17] to illustrate the distribution of *T* for each of the places described in the previous paragraph. These reveal the presence of associations between the place of early triage and the value of *T*. Statistical tests were carried out to help evaluate whether it is plausible that these associations occurred because of chance alone. For details of the matters outlined in this paragraph, please see Sections 2.2.5 and 3.2.

2.2.3 Early triage and the early admission of guilt

This sub-study was conducted to address the research question: 'Does early triage lead to early admission of guilt?'. *In this context*[18], early admission of guilt is taken to mean that, during an early interview, the suspect admits or partially admits the allegations, or admits to a lesser offence. As elsewhere in this report, the phrase early triage refers to triage that occurred during the scene attendance, or shortly afterwards, and before any early interview would have happened.

A positive triage finding is one that provides evidence of criminality. In the Safenet study as a whole, the only positive early triage findings that were known about were those which arose from the early triage of mobile phones, tablets or SIM cards. For this reason, *for the purposes of the Safenet study as a whole and for this sub-study*, a positive early triage finding is defined as one that provides evidence of criminality from one or more such devices and a negative early triage finding is one that does not.

In the Safenet dataset, each case is identified by a unique 'Anonymised Case Reference' and, as described in Section 2.1, all cases have at least seven rows in that dataset. This sub-study used the version of the Safenet dataset that is dated 14 May 2021.

To avoid artificially inflating the sample size (i.e., the number of rows) and to ensure that neither the under- nor the over-representation of cases occurred, it was necessary to use a subset of the Safenet data in this sub-study. To achieve this subsetting, a copy of the Safenet dataset was made and the row order in that copy was randomised. Then all rows were removed that contained duplicate values of all of 'Anonymised Case Reference', 'Date and time of scene attendance', and 'Date and time of early interview'. The rows of the Safenet dataset that remain after such a process will vary depending on the order of rows that is produced by the randomisation process. However, each such variant is a valid sample. This is because the criterion used to define a positive early triage

[17] For information on box and whisker plots, and their interpretation, see Galarnyk (2018)

[18] Elsewhere in this report, early admission of guilt means that, during one of the first two possible interviews, the suspect admits or partially admits the allegations, or admits to a lesser offence. Throughout this report, the term early interview refers to the first of these two possible interviews.

finding (i.e., that given above) and that used to define an early admission of guilt are each the same for all rows in the Safenet dataset.

After that subsetting process, irrelevant data were dropped by, in sequence, removing rows:

- for which there is no record of either the early triage findings or whether early triage occurred (this removed 2 rows);
- that contained an early interview that occurred before the scene attendance (this removed 1 row).

Then, to omit data that might introduce bias, those 40 of the remaining rows for which it was shown during the investigation that the suspect had *not* committed a crime were removed.

This resulted in a sample that contained 207 rows and 201 cases. Each of those rows represents an opportunity that early triage had to provide evidence that led to early admission of guilt in an instance in which it is believed that a crime was committed. The size of that sample is the number of rows that it contains. This is equal to, or very close to, the total number of such opportunities in the Safenet dataset.

To allow the analysis of the data that is presented in Section 3.3, the variables 'Early triage', 'Early interview', 'Early triage occurrence', 'Early triage finding' and 'Early admission' were generated from the sample referred to in the previous paragraph. These variables are defined as follows:

'Early triage'

The 'Early triage' variable is a factor (i.e., it contains categorical data) with three levels (i.e., it has three possible values). These levels are:

- 'Positive triage', meaning early triage occurred and *did* find evidence of crime on one or more mobile phones, tablets or SIM cards;
- 'Negative triage', meaning early triage occurred and did *not* find evidence of crime on one or more mobile phones, tablets or SIM cards;
- 'No triage', meaning early triage did not occur.

'Early interview'

The 'Early interview' variable is a factor with three levels. These are:

- 'Admission', meaning that, during early interview, the suspect *made* an early admission of guilt as defined above;
- 'Non-admission' meaning that, during early interview, the suspect did *not* make an early admission of guilt as defined above;
- 'No interview', meaning that there was no early interview and therefore the suspect had no opportunity to make an early admission of guilt as defined above.

'Early triage occurrence'

The 'Early triage occurrence' variable is a factor with two levels. These are:

- 'Triage occurred', meaning early triage happened;
- 'No triage', meaning early triage did not happen.

'Early triage finding'

The 'Early triage finding' variable is a factor with two levels. These are:

- 'Positive', meaning early triage happened and it *did* find evidence of crime on one or more mobile phones, tablets or SIM cards;
- 'Negative', meaning early triage happened and it did *not* find evidence of crime on one or more mobile phones, tablets or SIM cards.

'Early admission'

The 'Early admission' variable is a factor with two levels. These are:

- 'Admission', meaning that an early interview occurred, during which the suspect *made* an early admission of guilt as defined in the first paragraph of this Section (i.e., Section 2.2.3);
- 'Non-admission', meaning that an early interview occurred, during which the suspect did *not* make an early admission of guilt as defined in the paragraph referred to in the previous bullet point.

Analysis that makes use of these variables is given in Section 3.3. It contains contingency tables (also known as cross-tables)[19], bar charts and the use of statistical testing. See Section 2.2.5 for a rationale for the use of such testing and a description of how it has been used in this report.

A further variable, called 'Early interview occurrence', was also generated. For a definition of this variable and the analysis that it allowed, please see Appendix 8.

2.2.4 The time between referral and scene attendance, and the loss of digital evidence

This sub-study was conducted to address the research question: 'Does delay between the date of referral and the date of scene attendance lead to loss of digital evidence?'. It used the version of the Safenet dataset that is dated 14 May 2021. In this context, the referral means the receipt of the intelligence report that was sent to Staffordshire Police to alert them to the possibility of a case involving online child sexual exploitation.

The four indicators of digital evidence discovery that are detailed below were created to address this research question. Each of these is a factor (i.e., it is a variable that contains categorical data) with two levels (i.e., it has two possible values). For each indicator these levels are 'Yes' and 'No'.

Indicator 1: Digital evidence discovered during early triage?

For this indicator, 'Yes' means that digital evidence of criminality was found on one or more mobile phones, tablets or SIM cards during early triage and 'No' means that it was not.

Indicator 2: Sufficient evidence to proceed?

For this indicator, all rows were excluded from the subsample used that did not record that digital evidence had contributed to answering the 'Can we determine if a crime has been committed?' QP. Then, for this indicator:

- 'Yes' means that the final outcome of the case was *not* that it ended because of evidential difficulties;
- 'No' means that the final outcome of the case *was* that it ended because of evidential difficulties.

Indicator 3: Was there evidence that led to pre-charge admission of guilt?

For this indicator, 'Yes' means that digital evidence contributed to answering the "Was evidence provided that led to a suspect admitting their guilt prior to them being charged with the offence?" QP, and 'No' means that it did not.

[19] For an introduction to contingency tables, see Stover (n.d.).

Indicator 4: Was there evidence that directly led to referral for charging?

For this indicator, 'Yes' means that digital evidence contributed to answering the "Can evidence be provided that will directly lead to the referral of case for charging?" QP, and 'No' means that it did not.

As detailed in Table 2.5, a different subset of the Safenet database was used to generate each of the above-defined indicators. Then, from data held in each row of each of those subsets, the time that elapsed between the referral and the scene attendance was calculated, as was the value of the relevant indicator. For each indicator, box and whisker plots were then drawn[20] and a statistical test was completed to explore whether there is evidence of an association between the value of the indicator concerned and that elapsed time. These plots and tests are reported in Section 3.4. The sample size (i.e., number of rows) was 156, 194, 215 and 214 for indicators 1 to 4, respectively.

[20] For information on box and whisker plots, and their interpretation, see Galarnyk (2018)

Table 2.5 The subsets of the Safenet dataset that were used for each of indicators 1 to 4

Indicator		Subset	
Number	**Name**	**Initial subsetting process used**	**Rows then excluded from the initial subset prior to analysis***
1	Digital evidence discovered during early triage?	All rows of the dataset were randomised, then those rows that have duplicated values for all of "Anonymised Case Reference", "Date referral received" and "Date and time of scene attendance" were removed.†	Those which record that: • early triage did not happen; • the suspect had not committed any of the crime(s) they were suspected of committing; • the findings of early triage are not recorded; • scene attendance predates the receipt of the intelligence referral (this removed 2 rows).
2	Sufficient evidence to proceed?	Only those rows with the 'Establish crime committed' IP were selected.‡	Those rows which record that: • the suspect had not committed any of the crime(s) they were suspected of committing; • contain no record of a date of scene attendance; • scene attendance predates the receipt of the intelligence referral (this removed 3 rows); • digital evidence did not contribute to answering the 'Can we determine if a crime has been committed?' QP.
3	Was there evidence that led to pre-charge admission of guilt?	Only those rows with the 'Admission of guilt at an earlier stage' IP were selected.‡	Those rows which record that: • the suspect had not committed any of the crime(s) they were suspected of committing; • contain no record of a date of scene attendance; • scene attendance predates the receipt of the intelligence referral (this removed 3 rows).
4	Was there evidence that directly led to referral for charging?	Only those rows with the 'Referral for charging' IP were selected. ‡	Those rows which record that: • the suspect had not committed any of the crime(s) they were suspected of committing; • contain no record of a date of scene attendance; • scene attendance predates the receipt of the intelligence referral (this removed 3 rows).

* The intention of this row removal process was to remove both any irrelevant data points and those which may cause any signal to be unnecessarily lost in noise.
† The purpose of this procedure was to avoid the over or under-representation of cases in the subset that was analysed.
‡ The purpose of this procedure was to retain in the subset used for analysis only those rows which pertain to the relevant QP.

2.2.5 Statistical testing

The data contained in the complete Safenet dataset is a subset of all possible data of that type that could be recorded. The reason it is a subset is that its extent is restricted in various ways. For example, it only contains data recorded from cases that were referred to the police in a specific time period and in a specific geographical area. For each of the sub-studies detailed in Sections 2.2.1, 2.2.2, 2.2.3 and 2.2.4, to ensure that the data analysed was fit for purpose, it was necessary to use one or more subsets of that subset.

In statistical work, any given subset of all possible data is called a sample and all possible such data is known as the population. From the information given in the previous paragraph, it is clear that all of the analyses presented in this report have been conducted on samples. This opens the possibility that any patterns seen in those samples are not true reflections of patterns that are present in the population. It is possible, that by chance alone, the process that created the samples analysed has also created the patterns seen. One reason that this is important is because some patterns can be of operational interest. For example, imagine that analysis of a sample found that the choice to undertake a particular action by police was associated with a reduction in crime. If, on that basis, the use of that action was mandated, any expected reduction in crime would likely not happen if the pattern seen in the sample on which that policy change was based had occurred by chance alone. So, before making any operational change based on a pattern seen in a sample, it is wise to examine the plausibility that the pattern of interest has been caused by chance alone.

A common approach is to such an examination is to carry out a Null Hypothesis Significance Test (NHST)[21]. Many such tests have been devised for use in different situations. However, each of these results in a test statistic – which is given a symbol, such as *V*, *W* or χ^2, and a *p*-value. That *p*-value is the probability of obtaining the value of the test statistic that is seen, or one more extreme, if there is, in fact, no effect (pattern) in the population.

In statistical testing in general, and in the specific tests used in this study, an effect seen in a sample is considered to be statistically significant if its *p*-value is less than a critical value, α. Commonly, and in this study, α is set to 0.05. If *p* were to equal 0.05, there would be a 5% probability (i.e., a 1 in 20 chance) of the effect seen in the sample happening if there was no effect in the population. This is known as the Type 1 error rate. Traditionally, and in this study, the discovery that an effect seen in a sample is statistically significant is used to suggest that an effect at the population level has been detected. In other words, statistical significance is used to suggest that the effect seen in the sample has not happened by chance alone.

In some instances, in this study, multiple NHSTs have been carried out on the same data to explore the same research question. Under these circumstances, steps have to be taken to ensure that this does not inflate the Type 1 error rate[22]. In this study, to control this inflation, the Bonferroni correction (Field et al., 2012 p914) has been applied to the *p*-values as appropriate. On occasion, a situation will arise when the uncorrected *p*-value is < 0.05 but the corrected *p* is > 0.05. Under these circumstances, the effect would have been found to be statistically significant if it had been the only one tested; however, it is no longer found to be so once the *p*-value has been corrected to allow for the multiple NHSTs that have been carried out. This places the effect into to what can be called a discussible zone. Effects that fall into that zone cannot be considered to be truly significant. They can, however, be treated as interesting observations that warrants further work to retest the plausibility of the observed effect being due to chance alone. Importantly, the approach to such further work should *not* be the collection of more data to add to that already collected until statistical significance is seen. Such an approach would invite an extremely high chance of a wrongful conclusion that the observed effect was not caused by chance alone. Instead, work should be commenced that is specifically designed to rigorously test that possibility and to do so entirely on new data.

Statistical significance is not a measure of how large an impact an effect will have in the real world. Given large enough samples, very small effects (those with very little real-world impact) can be found to be statistically significant. Various measures of effect size have been devised for which Cohen (1988) has suggested benchmark values that represent small, medium and large effect sizes.

[21] For an excellent exposition of the principles and limitations of NHST, and related topics, see Spiegelhalter (2019).
[22] For a useful article on this, see Jafari and Ansari-Pour (2019).

Even when an effect seen in a sample is deemed not to have happened due to chance alone, this does not necessarily mean that the cause of that effect is known. This matter, in the context of the utility of the Safenet data, is explored in Section 4.1.2.

3 Results and Discussion

3.1 Dashboards from the Safenet dataset

As described in Section 2.2.1, Barrett and Varraich (2021) used the Safenet data to measure *Effectiveness* for each of the seven Impact Points (IPs) listed in Table 1.1 (in Section 1.5) and to estimate *Timeliness* for three of them (namely, Admission of guilt and an earlier stage', 'Establish crime committed' and 'Inform interview strategies'). They also used Excel pivot tables and pivot charts to produce interactive dashboards to display synoptic data drawn from that data. Those dashboards allow the observation and exploration of the *Effectiveness* and *Timeliness* of digital forensic science in addressing the Impact Points (IPs) referred to above. For each of those IPs, the dashboards also allow the data to be examined for the presence of associations between the values of each of *Effectiveness* and *Timeliness* and a number of factors, including the presence or absence of early triage[23].

Key results evident in these dashboards are given in Sections 3.1.1 and 3.1.2, and Section 3.1.3 provides an overview and commentary on those results.

3.1.1 *Effectiveness* when digital forensic methods were used

As described in Section 2.1, for each case in the Safenet dataset, there is one row for every opportunity that digital forensic methods had to contribute to the addressing of each of the seven IPs shown in Table 1.1. The frequency with which these contributions occurred in that dataset is a direct expression of *Effectiveness* as defined in Section 2.2.1.

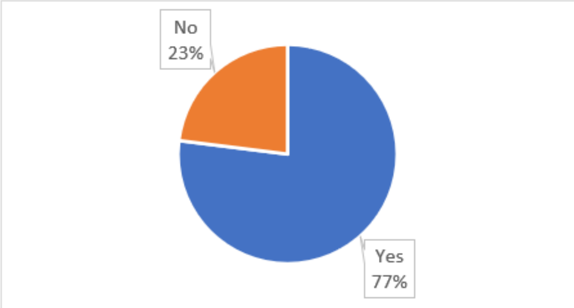
As illustrated by Figures 3.1 to 3.6, the dashboards provide convenient synopses of these *Effectiveness* data. The data used to create these Figures were drawn from 1258 rows (i.e., observations or Impact Point instances) of the Safenet dataset[24]. Those rows represent 239 cases that were notified to the police over the time period 06 July 2016 to 16 December 2019.

To create part (a) of each of Figures 3.1 and 3.2, the 1258 rows of interest were examined. From these, those 1129 rows were selected where the records show that digital forensics was used[25]. These represent 217 cases, which span the same date range as given above. The label 'Yes' was then given to those of the 1129 rows in which it was recorded that digital forensics made a contribution to addressing the IP that is stated in the row concerned. The remainder of those 1129 rows were labelled 'No'. These labels were then each counted and expressed as a percentage of 1129. The creation of part (b) for each of these two Figures was a directly analogous process; the only difference being that the label 'Yes' was only given to those 1129 rows where there was no record of any non-forensic method making a contribution to the addressing of the IP of interest. The difference between Figure 3.1 and 3.2 is that the latter shows the data grouped by IP.

[23] In this report, early triage means triage that commences during or soon after scene attendance and before any early interview would be expected to happen.
[24] For details of the data excluded prior to dashboard creation, see Table 2.2
[25] In this study, digital forensics were deemed to have been used only if at least one of the methods that was deployed during the investigation of the case concerned has the potential to be able to contribute to the answering of the relevant QP.

(a)

Did Forensics Contribute Overall?	Count	Percentage
Yes	869	76.97%
No	260	23.03%
Grand Total	**1129**	**100.00%**



(b)

Did Forensics Contribute Exclusively?	Count	Percentage
Yes	516	45.70%
No	613	54.30%
Grand Total	**1129**	**100.00%**

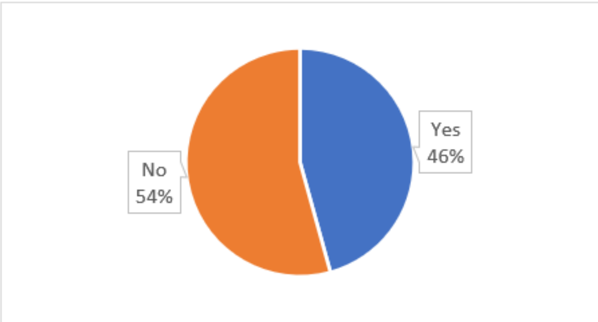
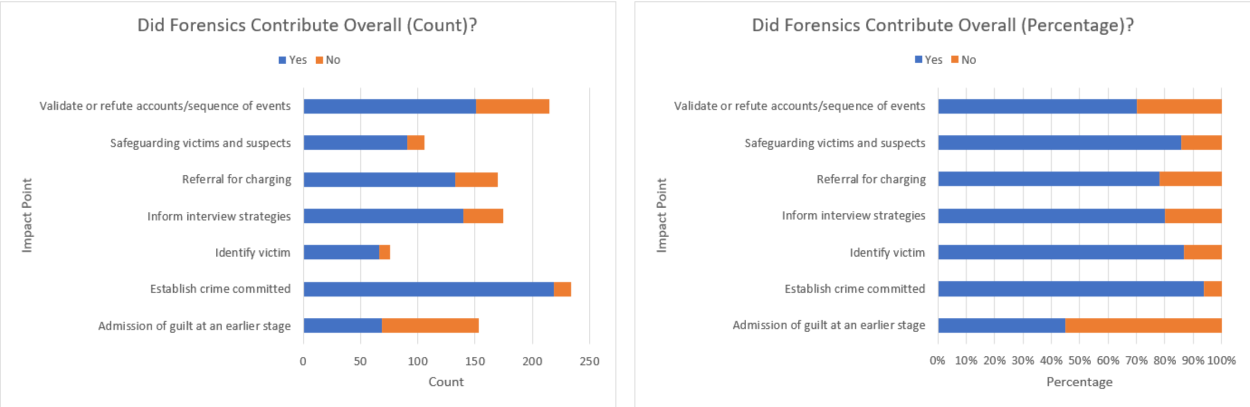


Figure 3.1 Dashboard screengrabs[26] that show, in cases where digital forensics was used, whether it made a contribution to the addressing of the seven IPs shown in Table 2.1. The category labelled 'Yes' in part (a) refers to observations where forensic methods made such a contribution, irrespective of whether non-forensic methods did so also. In contrast, in part (b), that category is reserved for observations where only forensic methods were involved in that contribution. These are direct expressions of the *Effectiveness* of the forensic methods used.

[26] This and all other screengrabs, and the data cited in this Section (i.e., Section 3.1), were taken from the spreadsheet named Staffordshire Analysis V2.xlsx as created by Barrett and Varraich (2021).

(a)

Did Forensics Contribute Overall?	Count		Percentage		Total Count	Total Percentage
	Yes	No	Yes	No		
Admission of guilt at an earlier stage	69	84	45%	55%	153	100%
Establish crime committed	219	15	94%	6%	234	100%
Identify victim	66	10	87%	13%	76	100%
Inform interview strategies	140	35	80%	20%	175	100%
Referral for charging	133	37	78%	22%	170	100%
Safeguarding victims and suspects	91	15	86%	14%	106	100%
Validate or refute accounts/sequence of events	151	64	70%	30%	215	100%
Grand Total	**869**	**260**	**77%**	**23%**	**1129**	**100%**



(b)

Did Forensics Contribute Exclusively?	Count		Percentage		Total Count	Total Percentage
	Yes	No	Yes	No		
Admission of guilt at an earlier stage	52	101	34%	66%	153	100%
Establish crime committed	142	92	61%	39%	234	100%
Identify victim	22	54	29%	71%	76	100%
Inform interview strategies	112	63	64%	36%	175	100%
Referral for charging	71	99	42%	58%	170	100%
Safeguarding victims and suspects	34	72	32%	68%	106	100%
Validate or refute accounts/sequence of events	83	132	39%	61%	215	100%
Grand Total	**516**	**613**	**46%**	**54%**	**1129**	**100%**

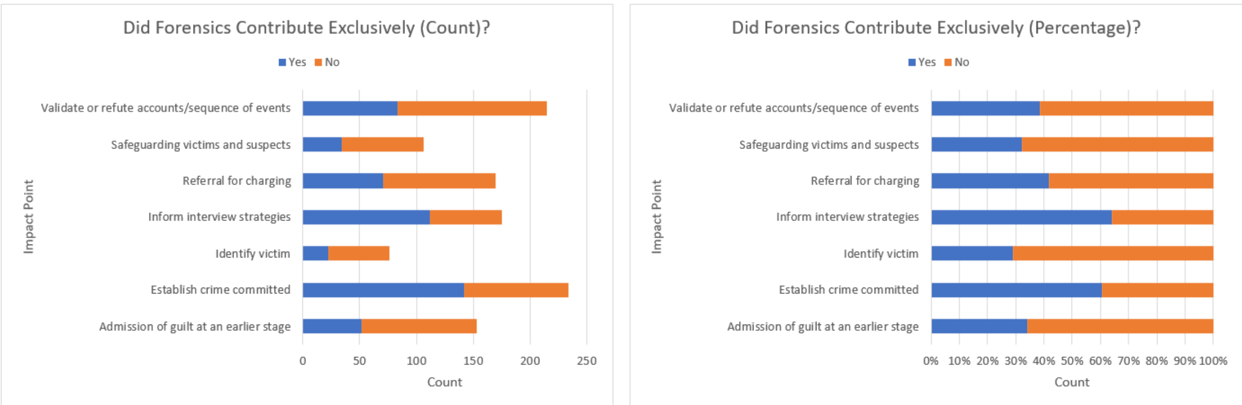
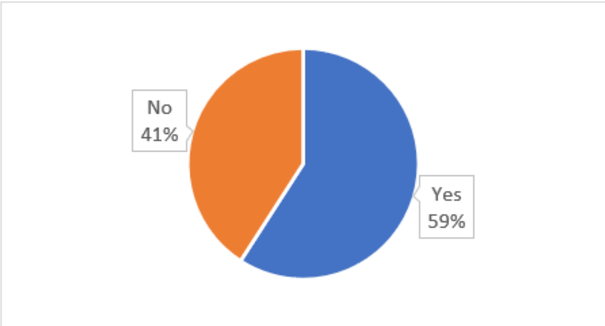


Figure 3.2 Dashboard screengrabs that show, in cases where digital forensics was used, whether it made a contribution to addressing each of the seven IPs shown in Table 2.1. The category labelled 'Yes' in part (a) refers to observations where forensic methods made such a contribution, irrespective of whether non-forensic methods did so also. In contrast, in part (b), that category is reserved for observations where only forensic methods were involved in that contribution. These are direct expressions of the *Effectiveness* of the forensic methods used.

Figures 3.3 and 3.4 are, respectively, the same as Figures 3.1 and 3.2, except that they only include those of the 1129 rows where records show that early triage was *not* used. There are 230 such rows, representing 51 cases. In contrast, Figures 3.5 and 3.6 are, respectively, the same as Figures 3.1 and 3.2, except that they only include those of the 1129 rows where records show that early triage *was* used. There are 899 such rows, representing 167 cases.

(a)

Did Forensics Contribute Overall?	Count	Percentage
Yes	136	59.13%
No	94	40.87%
Grand Total	**230**	**100.00%**



(b)

Did Forensics Contribute Exclusively?	Count	Percentage
Yes	76	33.04%
No	154	66.96%
Grand Total	**230**	**100.00%**

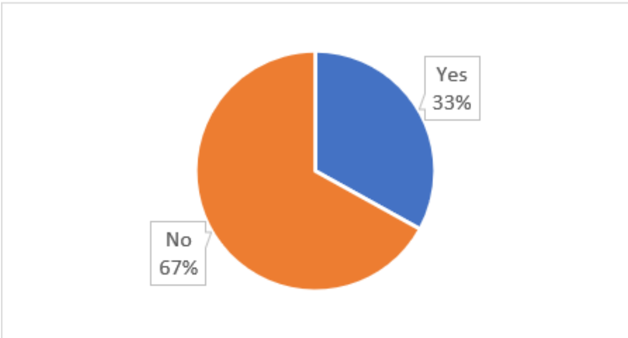
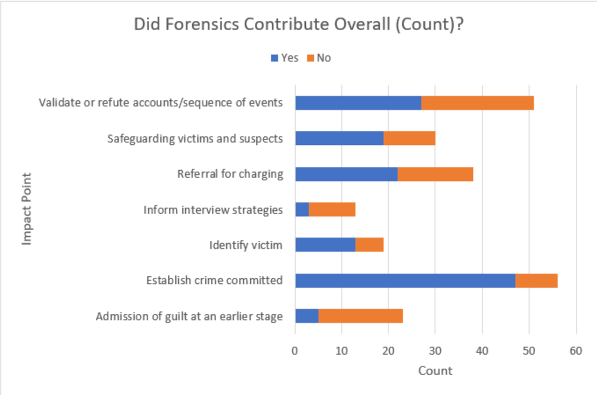
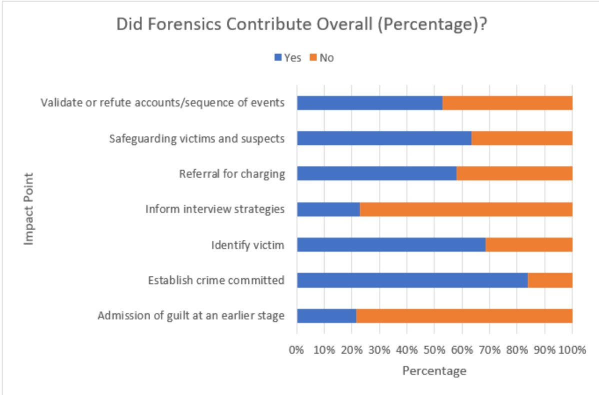


Figure 3.3 Dashboard screengrabs that show, in cases where digital forensics was used but early triage was not, whether forensics made a contribution to the addressing of the seven IPs shown in Table 2.1. The category labelled 'Yes' in part (a) refers to observations where forensic methods made such a contribution, irrespective of whether non-forensic methods did so also. In contrast, in part (b), that category is reserved for observations where only forensic methods were involved in that contribution. These are direct expressions of the *Effectiveness* of the forensic methods used.

(a)

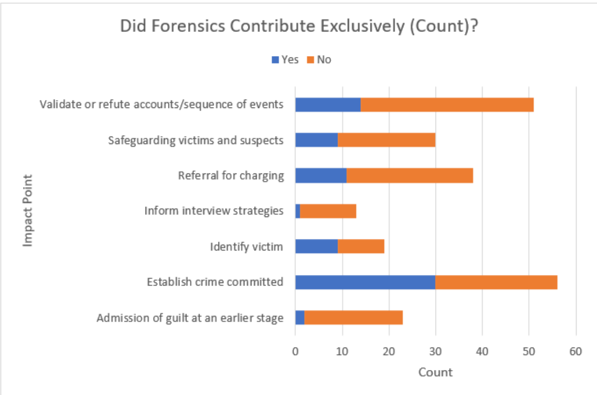
Did Forensics Contribute Overall?	Count Yes	No	Percentage Yes	No	Total Count	Total Percentage
Admission of guilt at an earlier stage	5	18	22%	78%	23	100%
Establish crime committed	47	9	84%	16%	56	100%
Identify victim	13	6	68%	32%	19	100%
Inform interview strategies	3	10	23%	77%	13	100%
Referral for charging	22	16	58%	42%	38	100%
Safeguarding victims and suspects	19	11	63%	37%	30	100%
Validate or refute accounts/sequence of events	27	24	53%	47%	51	100%
Grand Total	**136**	**94**	**59%**	**41%**	**230**	**100%**





(b)

Did Forensics Contribute Exclusively?	Count Yes	No	Percentage Yes	No	Total Count	Total Percentage
Admission of guilt at an earlier stage	2	21	9%	91%	23	100%
Establish crime committed	30	26	54%	46%	56	100%
Identify victim	9	10	47%	53%	19	100%
Inform interview strategies	1	12	8%	92%	13	100%
Referral for charging	11	27	29%	71%	38	100%
Safeguarding victims and suspects	9	21	30%	70%	30	100%
Validate or refute accounts/sequence of events	14	37	27%	73%	51	100%
Grand Total	**76**	**154**	**33%**	**67%**	**230**	**100%**



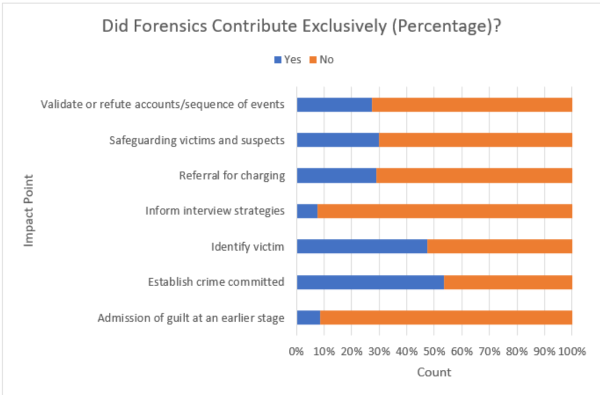
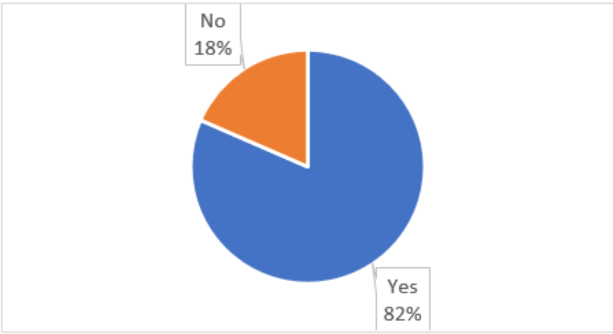


Figure 3.4 Dashboard screengrabs that show, in cases where digital forensics was used but early triage was not, whether forensics made a contribution to addressing each of the seven IPs shown in Table 2.1. The category labelled 'Yes' in part (a) refers to observations where forensic methods made such a contribution, irrespective of whether non-forensic methods did so also. In contrast, in part (b), that category is reserved for observations where only forensic methods were involved in that contribution. These are direct expressions of the *Effectiveness* of the forensic methods used.

(a)

Did Forensics Contribute Overall?	Count	Percentage
Yes	733	81.54%
No	166	18.46%
Grand Total	**899**	**100.00%**



(b)

Did Forensics Contribute Exclusively?	Count	Percentage
Yes	440	48.94%
No	459	51.06%
Grand Total	**899**	**100.00%**

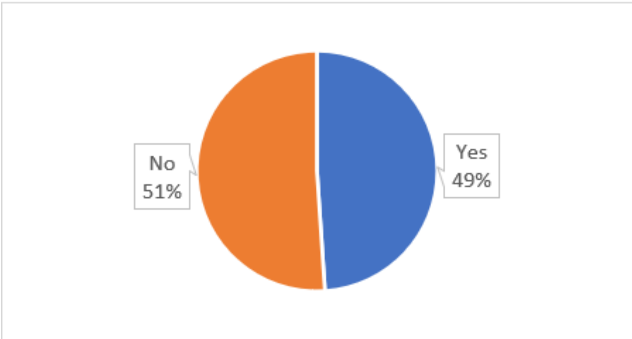
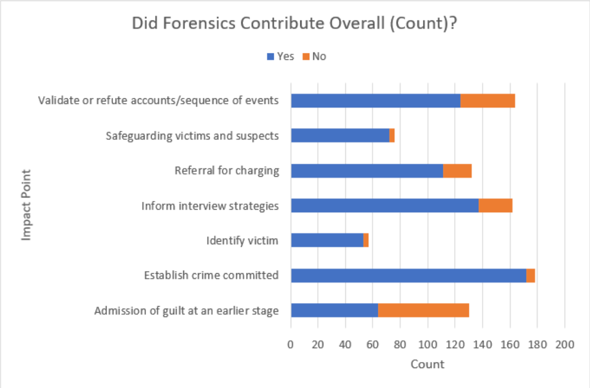
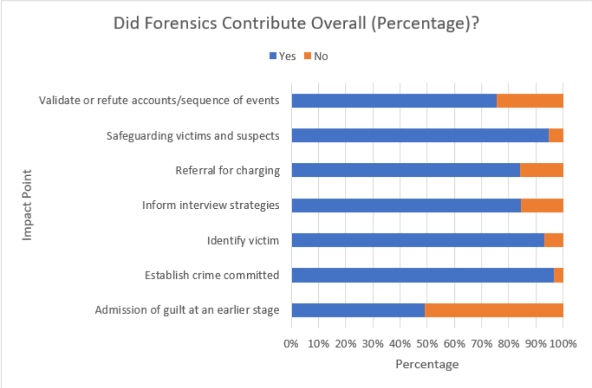


Figure 3.5 Dashboard screengrabs that show, in cases where both digital forensics and early triage were used, whether forensics made a contribution to the addressing of the seven IPs shown in Table 2.1. The category labelled 'Yes' in part (a) refers to observations where forensic methods made such a contribution, irrespective of whether non-forensic methods did so also. In contrast, in part (b), that category is reserved for observations where only forensic methods were involved in that contribution. These are direct expressions of the *Effectiveness* of the forensic methods used.

(a)

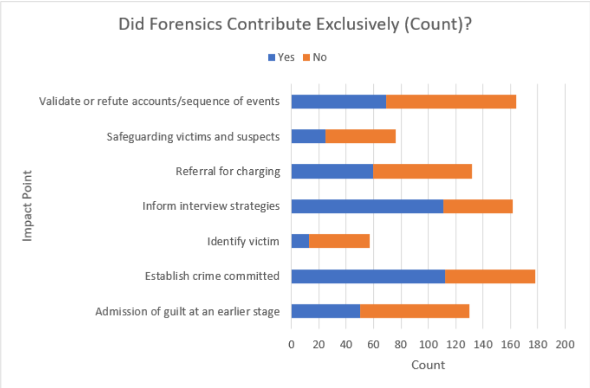
Did Forensics Contribute Overall?	Count Yes	No	Percentage Yes	No	Total Count	Total Percentage
Admission of guilt at an earlier stage	64	66	49%	51%	130	100%
Establish crime committed	172	6	97%	3%	178	100%
Identify victim	53	4	93%	7%	57	100%
Inform interview strategies	137	25	85%	15%	162	100%
Referral for charging	111	21	84%	16%	132	100%
Safeguarding victims and suspects	72	4	95%	5%	76	100%
Validate or refute accounts/sequence of events	124	40	76%	24%	164	100%
Grand Total	**733**	**166**	**82%**	**18%**	**899**	**100%**





(b)

Did Forensics Contribute Exclusively?	Count Yes	No	Percentage Yes	No	Total Count	Total Percentage
Admission of guilt at an earlier stage	50	80	38%	62%	130	100%
Establish crime committed	112	66	63%	37%	178	100%
Identify victim	13	44	23%	77%	57	100%
Inform interview strategies	111	51	69%	31%	162	100%
Referral for charging	60	72	45%	55%	132	100%
Safeguarding victims and suspects	25	51	33%	67%	76	100%
Validate or refute accounts/sequence of events	69	95	42%	58%	164	100%
Grand Total	**440**	**459**	**49%**	**51%**	**899**	**100%**



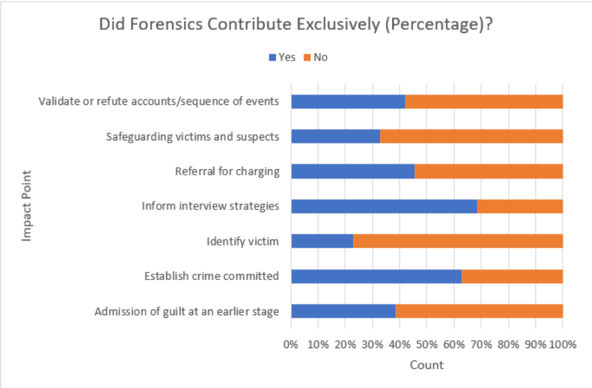


Figure 3.6 Dashboard screengrabs that show, in cases where both digital forensics and early triage were used, whether forensics made a contribution to addressing each of the seven IPs shown in Table 2.1. The category labelled 'Yes' in part (a) includes observations where both forensic and non-forensic methods made such a contribution. In contrast, in part (b), that category is reserved for observations where only forensic methods were involved in that contribution. These are direct expressions of the *Effectiveness* of the forensic methods used.

As can be seen from Figures 3.1 to 3.6, inclusive:

- when considering the 1129 rows of the Safenet data that represent observations when digital forensics was used and which were analysed to produce Figures 3.1 and 3.2, records show that:
    - 869 (77%) of these rows represent instances where digital forensics contributed to the addressing of an IP, and that for individual IPs, that percentage ranged from 45% (for 'Admission of guilt at an earlier stage') to 94% (for 'Establish crime committed');
    - 516 (46%) of these rows represent instances where digital forensics was the sole means by which an IP was addressed, and that for individual IPs, that percentage ranged from 29% (for 'Identify victim') to 64% (for 'Inform interview strategies')
- Figures 3.3 and 3.5 concern the analysis of rows in which it is recorded that digital forensics was used (for full details of the data excluded from this analysis, see the column headed 'Overview' in Table 2.2). Part (a) of each of these Figures show an association between the presence of early triage and the percentage of rows amongst in which it is recorded that digital forensics contributed to the addressing of an IP. When such triage was *not* used, that figure was 59%, which rose to 82% when it *was* used – a difference of 23 percentage points. Similarly, Part (b) of each of these Figures show an association between the presence of early triage and the percentage of rows amongst in which it is recorded that digital forensics made an exclusive contribution to the addressing of an IP. When such triage was not used, that figure was 33%, which rose to 49% when it was used – a difference of 16 percentage points.

- Figures 3.4 and 3.6 respectively concern the same data as used to build Figures 3.3 and 3.5, the difference being that in Figures 3.4 and 3.6, those data are grouped by Impact Point. Table 3.1 reproduces key aspects of the percentage data shown in those Figures that shows the grouped data.

Table 3.1 Percentage of rows of the data analysed in Figures 3.4 and 3.6 that show digital forensics contributing to the addressing of a specific Impact Point

Row No.[†]	Impact Point*	Nature of contribution**	Early triage used?	% of rows where contribution is recorded	Difference (in % points)[‡]
1	Admission of guilt at an earlier stage	Inclusive	No	22	27
			Yes	49	
		Exclusive	No	9	29
			Yes	38	
2	Establish crime committed	Inclusive	No	84	13
			Yes	97	
		Exclusive	No	54	9
			Yes	63	
3	Identify victim	Inclusive	No	68	25
			Yes	93	
		Exclusive	No	47	-24
			Yes	23	
4	Inform interview strategies	Inclusive	No	23	62
			Yes	85	
		Exclusive	No	8	61
			Yes	69	
5	Referral for charging	Inclusive	No	58	26
			Yes	84	
		Exclusive	No	29	16
			Yes	45	
6	Safeguarding victims and suspects	Inclusive	No	63	32
			Yes	95	
		Exclusive	No	30	3
			Yes	33	
7	Validate or refute accounts/ sequence of events***	Inclusive	No	53	23
			Yes	76	
		Exclusive	No	27	15
			Yes	42	

[†] These numbers are used later in this document to refer to specific rows in this Table.
* See Table 2.1 for the Questions Posed that correspond to these Impact Points
** An 'Inclusive' contribution is one in which digital forensics made a contribution *with or without* a contribution from non-forensic methods. An 'Exclusive' contribution is one in which digital forensics made a contribution *without* a contribution from non-forensic methods
*** This refers to the validation or refutation of the intelligence package received.
[‡] This is (% of rows where contribution and triage used is recorded) - (% of rows where contribution and triage not used is recorded), both rounded to the nearest whole number.

3.1.2 *Timeliness* when digital forensic methods were used

The analysis shown in Figures 3.7 to 3.9, inclusive, was carried out by examining 497 of the rows of the Safenet dataset, representing 215 cases that were all referred to the police between 6 July 2016 and 13 December 2019, inclusive. For details of the Safenet data excluded from that analysis, see the column headed '*Timeliness*' in Table 2.2. Figure 3.7 shows descriptive statistics drawn from all of those rows. In contrast Figures 3.8 and 3.9 each show such statistics drawn from different subsets of those rows. Those subsets are made up of those rows where it is recorded that either early triage was *not* used (Figure 3.8) or it *was* used (Figure 3.9).

Impact Point	Count	Minimum	Median	Maximum	Mean Average	Standard Deviation
Admission of guilt at an earlier stage	91	0	0	701	56.53	123.44
Establish crime committed	233	0	0	955	56.97	117.57
Inform interview strategies	173	0	0	557	46.94	102.32
Overall	**497**	**0**	**0**	**955**	**53.40**	**113.50**

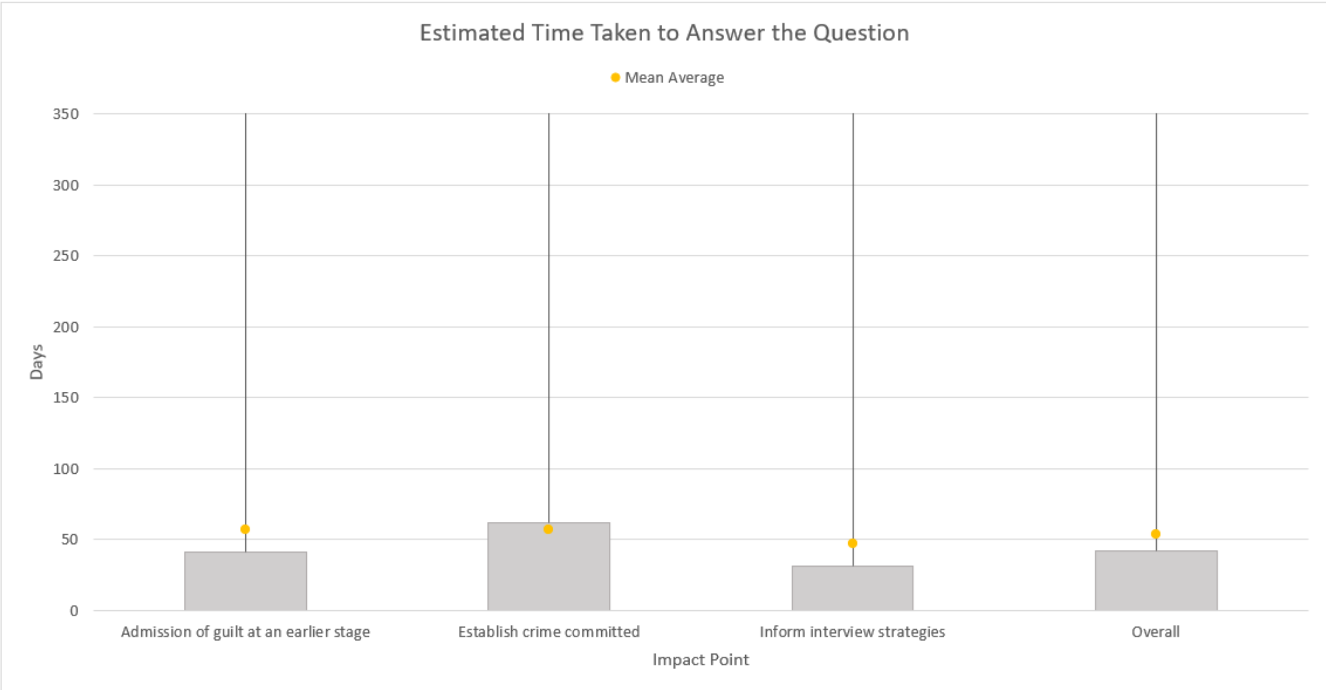


Figure 3.7 A dashboard screengrab that shows the estimated time taken to answer each of the Questions Posed that correspond with the three Impact Points given in the Table at the top of this screengrab (for the exact wording of these questions, please see Table 2.1). The data used in this analysis was restricted to 497 rows of the Safenet dataset in which it was recorded that digital forensics *was* used (see the column headed '*Timeliness*' in Table 2.2 for details of the data excluded).

Impact Point	Count	Minimum	Median	Maximum	Mean Average	Standard Deviation
Admission of guilt at an earlier stage	18	0	98.5	536	124.00	140.08
Establish crime committed	55	0	64	650	111.27	133.97
Inform interview strategies	13	0	64	557	108.92	149.61
Overall	**86**	**0**	**70**	**650**	**113.58**	**136.07**

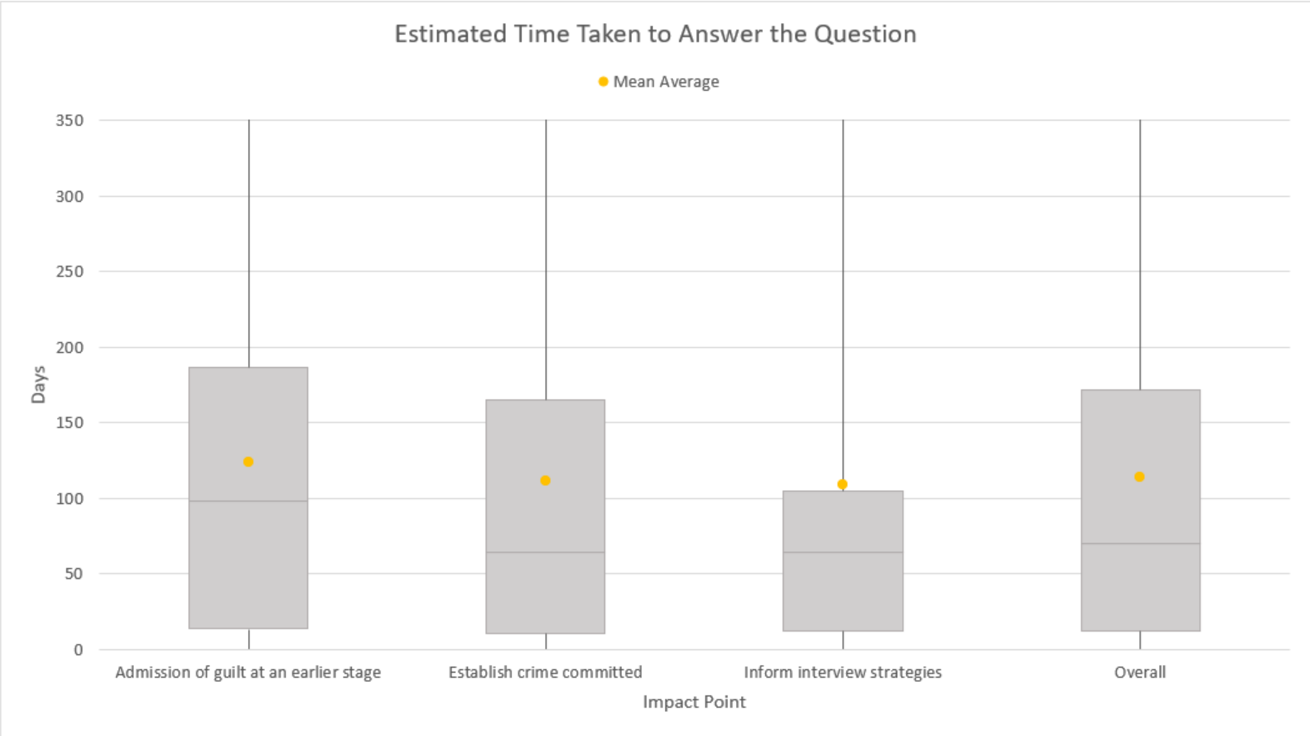


Figure 3.8 As Figure 3.7 but restricted to observations where early triage did *not* take place.

Impact Point	Count	Minimum	Median	Maximum	Mean Average	Standard Deviation
Admission of guilt at an earlier stage	73	0	0	701	39.89	113.99
Establish crime committed	178	0	0	955	40.20	106.98
Inform interview strategies	160	0	0	485	41.90	96.42
Overall	**411**	**0**	**0**	**955**	**40.81**	**104.06**

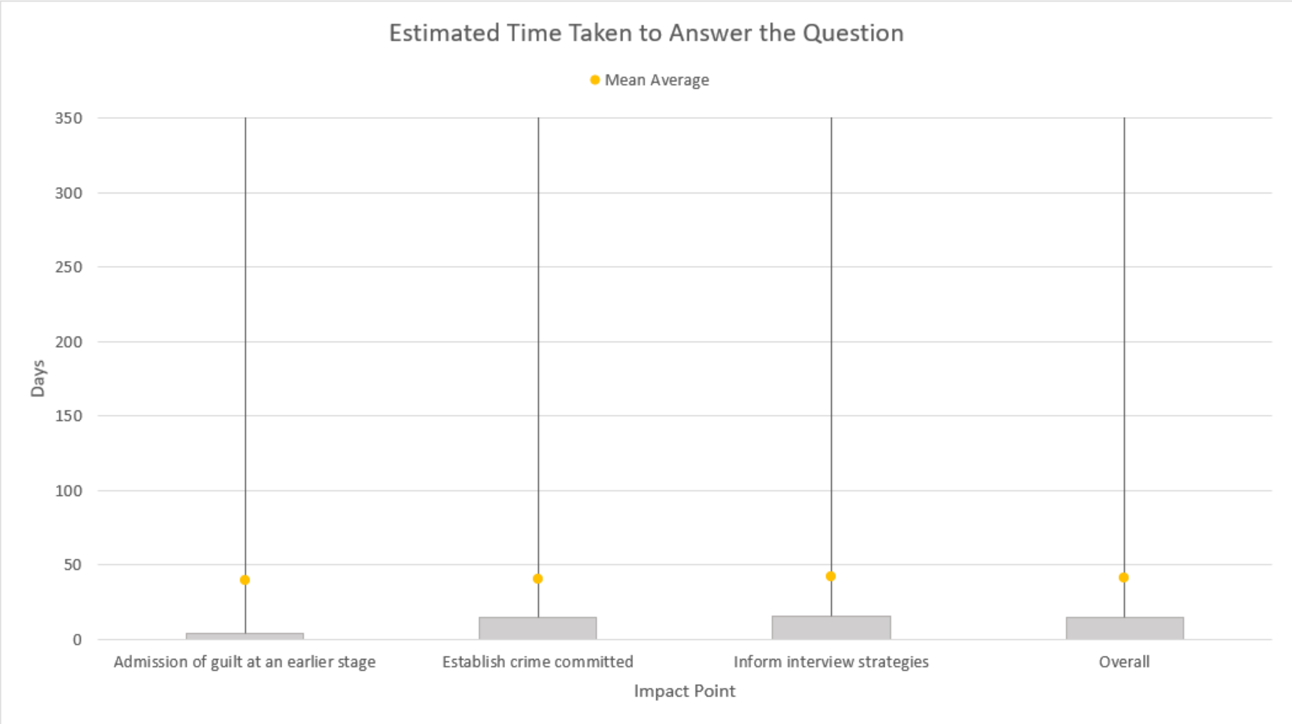


Figure 3.9 As Figure 3.7 but restricted to observations where early triage *did* take place.

Table 3.2 shows the three IPs in question, their corresponding QPs, the relevant typical (median) values taken from Figures 3.8 and 3.9 and the relevant differences between those medians. These medians are expressions of the metric *Timeliness* as defined in Section 2.2.1.

It is evident from Table 3.2 that, amongst the Safenet data analysed, the introduction of early triage is associated with very noticeable improvements in *Timeliness*. Such improvement is seen for each of the three Questions Posed for which that metric has been calculated.

Table 3.2 *Timeliness* as expressed by the median estimated times to answer the Question Posed based on the analysis of data held in 497 rows of the Safenet dataset. In each of these rows it was recorded that digital forensic methods were used. (See the column headed '*Timeliness*' in Table 2.2 for details of the data excluded).

Row No.[†]	**Impact Point (IP)**	**Question Posed (QP)**	**Median time *without* early triage/day**	**Median time *with* early triage/day**	**Time difference /days**
1	Admission of guilt at an earlier stage	Was evidence provided that led to a suspect admitting their guilt prior to them being charged with the offence?	98.5	0	98.5
2	Establish crime committed	Can we determine if a crime has been committed?	64	0	64
3	Inform interview strategies	Can we determine information that will inform our interview strategy?	64	0	64

[†] These numbers are used later in this document to refer to specific rows in this Table.

3.1.3 Overview and commentary

Effectiveness, Timeliness and associations seen

Consider the data that are presented in Figures 3.1 to 3.6, inclusive, and in Table 3.1 – all of which provide direct measures of the *Effectiveness* of the forensic methods used. Those data show that, amongst the observations analysed to produce them and when digital forensic methods were used:

- those methods frequently made a contribution to the addressing of each of the seven IPs of the Safenet study and that, in many instances, such methods were the sole contributor to this process;
- when early triage was used:
  - for all seven IPs, there was an increase in the percentage of rows in which it is recorded that digital forensic methods made a contribution to the addressing of the IP of interest;
  - for six of the seven IPs, there was an increase in the percentage of rows in which it is recorded that digital forensic methods made an exclusive contribution to the addressing of the IP of interest.

Also consider the data presented in Figures 3.7, 3.8 and 3.9, and in Table 3.2 – all of which provide direct measures of the *Timeliness* of the forensic methods used. Of these, Figures 3.8 and 3.9, and Table 3.2 show the apparent effect of the presence or absence of early triage on *Timeliness*. These show that for the three IPs under consideration that the typical (median) estimated time taken to answer the corresponding QP dropped from either 64 or 98.5 to 0 days on the introduction of early triage.

Explanation of Effectiveness

Figures 3.1 and 3.2 clearly illustrate the contributions made by digital forensic methods to the attainment the seven IPs shown in Table 2.1 – this is a direct measure of the *Effectiveness* of these methods. These contributions are explained by the ability of those methods to provide information of direct utility in answering the seven questions, also given in Table 2.1, that underpin those IPs.

Those Figures also show variation from case to case and from IP to IP in the utility of digital forensic methods in the attainment of those IPs. As discussed below, some of this variation is, at least in part, accounted for by the presence or absence of early triage.

Explanations of associations seen

There are three possible explanations of the association between the use of early triage and the performance enhancements seen in Figures 3.1 to 3.9 and Tables 3.1 and 3.2. These are that:

1. they happened by chance;
2. there is a common cause of both the use of early triage and the enhancements seen;
3. they are caused by the use of early triage.

Let us explore each of these in turn.

Explanation 1: Chance

If chance alone were responsible for the effects of early triage seen in the Safenet data (the sample), this would mean that these effects would not be present in the population (i.e., all possible cases of the type investigated by Operation Safenet). Arguably, given the relatively large size of both the overall sample and the effects seen, it would seem likely that the probability of this would be small. Formal statistical testing could be carried out to quantify this probability. To illustrate this, the exact Wilcoxon signed-rank test has been applied to the data shown in Table 2 of the Executive Summary. This revealed that the effect seen in that Table is statistically significant ($V = 28$ and $p = 0.016$) and so is not likely to have occurred by chance alone.

Explanation 2: Common cause

As a matter of policy, whenever suitably qualified personnel from Staffordshire Police's Digital Forensics Unit (DFU) are available at the time of scene attendance, such staff are deployed to conduct early triage in Operation Safenet cases. Importantly, this policy was in place throughout the period when the Safenet data were generated. This opens the possibility that fluctuations in the workload of the DFU caused variation in the availability of such staff, causing alterations in the occurrence of early triage[27]. If this mechanism were at play during the generation of the Safenet data, times of high workload in the DFU would coincide with periods of low probability of early triage use.

[27] This hypothetical causative link would not be possible if police officers routinely carry out early triage in the absence of DFU staff. We believe that there is good evidence to suggest that this does not happen. There are 632 rows of the Safenet dataset (the version dated 14 May 2021) in which it is recorded that the deployment of DFU staff to carry out early triage did *not* happen. Of these 632 rows, 563 (89.1%) show that early triage did *not* occur. In contrast of the 1336 rows that record that such deployment of DFU staff *did* happen, only 7 (0.5%) show the absence of early triage. The chi-square test (with Yates' continuity correction) of association between such DFU staff deployment and the occurrence of early tirage shows this effect to be highly significant ($\chi^2 = 1631, p < 2.2 \times 10^{-6}$). This strongly suggests that it has not happened by chance alone. Based on this evidence, and the fact that the purpose of the afore-mentioned DFU staff deployment is to conduct early triage, we contend: 1. that such deployment is instrumental in the undertaking of such triage and that 2., whilst early triage can be undertaken by police officers, it is unlikely to occur in any given Safenet case – at least in the absence of DFU staff deployed to carry out such work.

It may be possible that high workload in the DFU led not only to fewer instances of early triage but also to a more targeted approach to the collection of items during scene attendance. If this happened, this could lead to fewer items being retrieved at times of high DFU workload. This would likely lower the amount of available digital evidence, thereby decreasing the probability that digital forensic methods could contribute to IP attainment. Therefore, under this scenario, increased workload in the DFU would provide a common cause for decreases in the probability of both early triage occurrence and the DFU providing evidence that would contribute to IP attainment. This common cause would explain the effect seen in Table 3.1.

Consider a scenario in which high workload in the DFU prolongs the typical time taken between the receipt of an item by the DFU and its analysis. If such workload also decreases the opportunity for early triage as hypothesised above, fluctuations in DFU workload would also provide a common cause that would explain the associations see in Table 3.2.

Crucially, the common cause mechanisms outlined above are predicated on fluctuations in demand affecting the workload of the DFU. There is reason to believe that this is *not* what happened in the past or what happens now. Operational experience (Millar, 2021) suggests that demand for the services of the DFU always outstrips its capacity, implying that its workload does *not* vary with changes in that demand. Also, from that experience it is known that, for the period when the Safenet data were generated, DFU staff were always deployed to undertake early triage except when specific circumstances prevailed. These were when the DFU was provided with insufficient notice of scene attendance and either multiple staff were on leave or that attendance was planned for a weekend. This operational knowledge does *not* rule out the possibility of common cause explanations but it does throw substantial doubt on those outlined above.

On page 54, the reader is invited to note seven points in connection with the effects seen in Table 3.2. One of those (point 7) may have common cause explanations. These are explored on pages 54 and 55.

Further discussion of the utility of the Safenet data in the understanding of cause and effect is given in Section 4.1.2.

Explanation 3: Early triage use causes the effects seen

First, let us consider the effects seen in Table 3.1.

There are 1126 rows of the Safenet dataset[28] that show that digital forensic methods were used. Of these, there are 294 (i.e., 26%) in which it is recorded that:

- items were submitted to the DFU for analysis and
- there is only one interview and
- that sole interview occurred between scene attendance and the submission of items to the DFU for analysis.

Under these circumstances, in the absence of early triage, the interview would have occurred before there was any possibility of digital forensic evidence being available to inform interview strategies or lead the suspect to admit guilt prior to charging (admission of guilt can only happen during an interview). As early triage could provide such evidence, its use in these circumstances could lead to the addressing of the IPs in rows 4 and 1 of Table 3.1. This gives a plausible mechanism that could account for at least a proportion of each of the associations evident in that Table which concern

[28] Version dated 14 May 2021.

those IPs. However, that mechanism would not explain any of the associations seen in that Table for the other five IPs of the Safenet study.

In Operation Safenet, one of the purposes of early triage is to identify digital devices that show evidence of criminality, and those that do not[29]. It may be possible to quickly eliminate items from the investigation that fall into the second category[30]. When this is possible, it has the benefit of allowing their early return to their owners. This may serve to decrease or eliminate unnecessary inconvenience and worry for innocent people who might otherwise fear being suspects in the case.

Due to its ability to prioritise digital devices for further analysis, triage has been seen as a solution to the large amount of data that needs to be processed in digital forensic laboratories (Shaw and Browne, 2013). With good reason, then, it was believed that the afore-mentioned return of devices to their owners would also have the operational benefit of reducing the number of items that would be submitted to the DFU for analysis. This Safenet study has provided the opportunity to test that belief and, surprisingly, it was found to be incorrect. Indeed, there is some evidence to suggest that, the number of items submitted to the DFU for analysis might typically be greater when early triage is used than when it is not[31]. Furthermore, when early triage is used, it is known in advance of that analysis that there is evidence of criminal activity on those items – which is not the case for items submitted without such triage. With all of this in mind it is reasonable to hypothesise that the number of items that provide evidence of criminality on analysis in the DFU is typically greater when early triage is used than when it is not. If true, this provides a mechanism that can explain at least some of the performance improvement that is associated with the use of early triage that is evident in Table 3.1.

Yet, that mechanism does not address the reason why the use of early triage is not associated with a decrease in the typical number of items submitted for analysis by the DFU. One possibility has already been explored. This is that high demand in the DFU leads to a decrease in the probability of both the use of early triage and the recovery of any given item of evidence found at the scene. Under this possibility, the latter effect would be caused by a more targeted approach to exhibit collection instigated to manage that demand. However, there is no policy in place for such an approach to demand management – so if it has occurred it is solely because of the initiative of individual police personnel. Also, for this possibility to be a reality, there must be changes in DFU workload caused by alteration in the demand for its services; for reasons set out previously, this seems improbable.

It is noteworthy that, in the absence of early triage, the decision concerning which items to seize becomes solely a matter of police discretion. Under these circumstances, police officers would be expected to decide not to seize items that offer little *prima facie* reason to suspect that they contain evidence of illegal activity. However, in the absence of early triage, such decisions may have to be

---

[29] Information obtained from early triage is also used: 1. to attribute digital devices to suspects, 2. during interview and 3. for disclosure to the defence (Hood 2021).

[30] There are limitations to the capabilities of early triage and, consequently, devices that do not provide evidence of criminal activity during such triage may nonetheless be sent for full digital forensic analysis (Hood 2021).

[31] Amongst the data analysed, the median number of items that were submitted to the DFU for analysis was 2 in the absence of early triage and 3 in its presence. This difference of 1 is just statistically significant (Wilcoxon rank sum test with continuity correction reveals W = 6712 and p = 0.013). The removal of cases that were deemed to be not in the public interest from the data analysed reduces the difference between medians from 1 to 0.5 and that reduced difference is no longer statistically significant (Wilcoxon rank sum test with continuity correction reveals W = 3471 and p = 0.83). However, it is interesting to note that this analysis provides no evidence to suggest that the number of items submitted to the DFU for analysis is reduced by the deployment of early triage.

taken on the basis of little concrete information[32] and so, although made in good faith, some may be wrong. It is also possible that the police officer(s) attending a scene may not have the experience in identifying more unusual and/or disguised digital devices which would be identified as potential evidence by DFU personnel. This provides alternative mechanisms to explain why early triage is not associated with the recovery of fewer items. As this alternative is in keeping with the strictures of policy, and it is not dependent on fluctuations in the workload of the Digital Forensics Unit (DFU), it is perhaps more plausible than the explanation provided in the preceding paragraph.

During the period when the Safenet data were generated, in cases where early triage occurred, it was the norm for staff from the DFU to have been deployed to undertake that task[33]. It is possible, therefore, that specialist advice from such staff during such triage might impact on the typical number of items recovered for later analysis in the DFU. This could also provide possible reasons why early triage is not associated with the recovery of fewer such items. Included amongst these reasons are the possibility that specialist advice during early triage might improve the identification of potential exhibits. Also, such advice might result in the more frequent separation of devices and peripheries prior to submission, which would tend to increase the number of items submitted to the DFU for later analysis.

There is a further causative mechanism that deserves mention here, although its occurrence seems likely to be rare. Consider a scenario in which staff attend a scene to conduct early triage and they observe IIOCs recovered from items found at that scene. It is feasible that they could connect those images to that scene because features of the scene (e.g., wallpaper patterns) are apparent in those images. This connection could provide information that could contribute to the realisation of one or more of the seven IPs and so could explain at least some of the effects seen in Table 3.1. However, it cannot be known from the Safenet data how often, if at all, this scenario actually occurred.

Also, as explored in Section 3.3, the discovery of evidence of criminality during early triage is associated with increased likelihood of early interview occurrence. This is important because admission of guilt can only happen during an interview. It seems probable that this is a key driver behind the improvement in *Effectiveness* for the 'Admission of guilt at an earlier stage' IP that is associated with the use of early triage amongst the Safenet data analysed to produce Table 3.1

From the above it can be seen that there are multiple possible mechanisms by which the use of early triage could affect the improvements in *Effectiveness* that are evident in Table 3.1. Whilst there is evidence from the Safenet data that support some of these, all of them must remain at least largely hypothetical until further work can be completed to test them.

[32] To aid in this decision-making process, police officers might also be able to use information on the likely owner of a given device. Also, they may be able to scroll through the messages or other items held on a device (provided that they are not hidden or deleted). Such scrolling activity *would* fall within the definition of early triage as used in this study, provided that it was undertaken during scene attendance or shortly afterwards and before any early interview would be expected to occur.

[33] There are 1398 observations in the Safenet dataset (the version dated 14 May 2021) in which it is recorded that early triage occurred. DFU staff were deployed to carry out that tirage in 1329 (95.1%) of these.

Now let us consider the effects seen in Table 3.2.

Please note the following seven points:

1. The data in Table 3.2 solely concerns cases in which forensic methods were used[34].
2. Appendix 7 gives the decision trees[35] that were devised and used by Barrett and Varraich (2021) to estimate the time taken to answer each of the questions that appear in that Table. Following the branches of those trees for when forensics was used shows that whenever:
   - early triage *was* used, the:
     - start date used was the date of scene attendance;
     - end date used *did* include either:
       - the date of scene attendance (when the Question Posed was 'Can we determine if a crime has been committed?) or
       - the date of early interview (when the Question Posed was 'Was evidence provided that led to a suspect admitting their guilt prior to them being charged with the offence?' or 'Can we determine information that will inform our interview strategy?');
   - early triage was *not* used, the:
     - start date used was the date of DFU submission;
     - end date must be on or after the work of the DFU has started.
3. When early triage was used, it occurred on the same day as scene attendance.
4. When triage detected evidence of criminality it did so essentially instantaneously. This is true, whether that triage was early or part of the work done after DFU submission.
5. When early interviews occurred, they frequently happened on the day of scene attendance (85.9% of the relevant rows of the Safenet dataset that record the occurrence of both early triage and early interviews show that these interviews happened on the same date as did scene attendance).
6. DFU work can start on the day of DFU submission. However, this was not the norm. Amongst the relevant rows of the Safenet dataset which record that items were submitted to the DFU for examination, 63.3% show DFU work starting after the date of that submission.
7. Amongst the relevant rows of the Safenet dataset which record that items *were* submitted to the DFU for examination, whether DFU work started on or after the date of that submission varies with whether early triage was used. When early triage was *not* used, 17.4% of these rows show DFU work starting on the same date as that submission. That figure rises to 43.5% when early triage *was* used. This association between the occurrence of early triage and when DFU work started is statistically highly significant[36], indicating that it has not happened by chance alone.

Let us first consider possible explanations for the association reported in point 7. That association is interesting because it could be seen to suggest that perhaps items that were subjected to early triage are given a higher priority for DFU work than those which were not. If this were the case, this

---

[34] In the context of the Safenet study, the terms 'forensic methods', 'digital forensic methods' and 'forensics' are synonymous. Also, in the context of that study, forensic methods are deemed to have been used only if the method concerned both 1. has the potential to be able to contribute to the answering of the relevant QP and 2. was actually deployed to address that question during the investigation of the case concerned.

[35] In Appendix 7, the triage it refers to is early triage. Also, in the Safenet study as a whole (including in that Appendix), the variable 'Forensics Used' only has the value 'Yes' if, for at least one of the forensic methods employed, it both 1. has the potential to be able to contribute to the answering of the relevant QP and 2. was actually deployed to address that question during the investigation of the case concerned..

[36] The chi-square test of association with Yates' continuity correction gives $\chi^2 = 31.5, p = 1.99 \times 10^{-8}$.

would provide a causative link between early triage and that association. However, there is no policy in place for such priority to be given because early triage provides evidence that allows the risk attached to a given case to be better understood than for cases where no triage has been undertaken (Millar 2021). There is second possible explanation that is plausible at first glance. *If* early triage were *not* used at times when the services of the DFU were in exceptionally high demand, it is likely that more untriaged items would be received by the DFU during such periods. It might be expected that, at such times, pressure of work would also make it less likely that the examination of any given item would start on the day of submission. If true this would, in part at least, account for the association in question. However, as pointed out previously, the operational experience is that DFU capacity is always exceeded by demand (Millar 2021). This means that the decision not to proceed with early triage is not a consequence of unusually high demand. This would seem to exclude this second possibility as a plausible explanation. There is a third possible mechanism that merits attention, however. During the period of the Safenet data generation, the only occasions on which DFU staff were not deployed to carry out early triage happened when the DFU was provided with insufficient notice of scene attendance and either multiple staff were on leave or that attendance was planned for a weekend (Millar 2021). Also, during that period, early triage rarely took place in the absence of DFU staff deployed to carry it out[37]. Consider what would happen *if* high levels of DFU staff absence due to leave meant that both early triage was unlikely to occur and work on any given item received by the DFU was less likely to start on the day of submission. Under these circumstances the absence of multiple staff due to leave would be a common cause for both the absence of early triage and delays in the commencement of work on items submitted to the DFU. This would explain, at least in part, the association set out in point 7. If it were the only mechanism at play, it would mean that there is no causative link between early triage and that association. In turn it would mean that any of the improvement in *Timeliness* on the use of early triage that is due to that association would not be caused by the use of such triage.

Now, let us turn our attention to points 1 to 6. Whenever the start and end dates referred to in point 2 are the same, the estimated time to answer the Question Posed is 0 days. The Question Posed given in row 2 of Table 3.2 is ‘Can we determine if a crime has been committed?’. From points 2, 3, 4 and 6 it is clear that, for this Question Posed, there was always the opportunity for these dates to coincide when early triage was used but this is not so when early triage was not employed. Respectively, the Questions Posed given in rows 1 and 3 of Table 3.2 are ‘Was evidence provided that led to a suspect admitting their guilt prior to them being charged with the offence?’ and ‘Can we determine information that will inform our interview strategy?’. Points 2, 3, 4, 5 and 6 indicate that the frequency of the possibility for answering these questions on the start date would be higher when early triage was deployed than when it was not. These observations provide plausible mechanisms by which explanation 3 (i.e., that early triage use causes the effects seen) could occur.

It is clear that there are plausible mechanisms by which early triage could cause the beneficial effects seen in Tables 3.1 and 3.2. These mechanisms have a common thread. This is the ability of early triage to uncover evidence of criminality at a very early point in the processing of cases of online CSE. This means that those beneficial effects can be well explained by the ability of early triage to inform the decision making of key personnel during the crucial early stages of an investigation. However, before this explanation can be accepted, further work would be needed to exclude the possibility of one or more common cause explanations.

[37] There are 1398 observations (i.e., rows) of the version of the Safenet dataset dated 14 May 2021 in which it is recorded that early triage occurred. Of these, only 69 (4.1%) show that this work happened in the absence of DFU staff deployed to carry out this task.

Cognitive bias

Cognitive bias is a collective term for systematic errors of thought, each of which – whether conscious or not – cause a skew in the data recorded or the understanding drawn from them. Any given study that is affected by cognitive bias will have a predisposition towards a particular conclusion (Jackson and Jackson 2017 p5).

All of the data held within the Safenet dataset, that was drawn without interpretation from case files, are from cases that happened before the Safenet study started. One of the advantages of this approach is that that it excludes the possibility that any form of cognitive bias that could have been caused by the introduction of that study.

However, the generation of some of the data contained in that dataset did necessitate the interpretation of information held on casefiles. For the reasons given below, those data could have been affected by contextual bias, which is one of several forms of cognitive bias that have been recognised.

Contextual bias happens when reasoning and/or judgement is skewed by the introduction of extraneous information, predisposing the outcome of those cognitive processes in a particular direction. The data capture process that created the Safenet dataset is described in Section 2.1. That work required researchers to decide which of any forensic and non-forensic methods used in a particular case contributed to the answering of each of the seven questions listed in Table 2.1. This necessitated those researchers exercising of both reasoning and judgement. Throughout the data capture process, those researchers were aware of the aims of the Safenet study. Specifically, they knew that it is aimed at the measurement of the impact of digital forensic methods amongst the cases of interest and the effect of early triage on that impact. Therefore, if it is accepted that this knowledge is extraneous to their roles, it could have produced unconscious contextual bias.

If present, that bias could have produced a number of misleading effects. For example, it could have exaggerated the enhancement in the *Effectiveness* of digital forensics that is associated with the use of early triage that is apparent in Table 3.1. Fortunately, each row of the Safenet dataset contains a column in which the final outcome of the investigation in question is recorded. The entries in that column are taken directly from casefiles and so require no reasoning or judgement. This means that they are not susceptible to the effects of any unconscious contextual bias that might be present. Those entries each take the form of one of the eleven outcomes shown in Table 3.3. A key aim of an investigation is to establish whether or not the suspect committed a crime. Table 3.3 also classifies those outcomes into groups dependent on whether they show that that aim has been achieved.

Table 3.3 The possible final outcomes of an investigation as recorded in the Safenet dataset, classified according to whether each such outcome shows whether the investigation established whether or not the suspect committed a crime.

Outcome	Did the investigation establish whether or not the suspect committed a crime?
Awaiting sentencing	Yes
Charged	Yes
Community resolution	Yes
Conditional caution	Yes
Youth conditional caution	Yes
Deported	Yes
Not in the public interest (i.e., the suspect had not committed a crime)	Yes
Evidential difficulties	No
NFA	No
Ongoing investigation	Not known (as the investigation was incomplete at the time that the Safenet data were recorded)
Deceased	Not known (as the suspect died before the investigation was complete)

A subset of the Safenet dataset was created by:

- excluding those rows in which it is recorded that the final outcome of the investigation was either 'Ongoing investigation' or 'Deceased' then
- keeping only those rows in which it is recorded that digital forensic methods were used.

Amongst the rows of that subset that show that early triage was *not* used, 67.0% show that the investigation established whether or not the suspect committed a crime. This figure rises to 82.4% when early triage *was* used. This association is statistically highly significant[38], indicating that it has not happened by chance alone.

The association set out in the previous paragraph is entirely consistent with the results shown in Table 3.1. This does not prove that the data are free of unconscious contextual bias. However, it provides some degree of reassurance that, if such bias is present, its effects do not seem to be extreme.

Conclusion

As shown in Figures 3.1 and 3.2, amongst the cases captured in the Safenet dataset, when digital forensic methods were used those methods made a substantial contribution to the realisation of the seven IP given in Table 2.1. As evidenced by the variation seen in those Figures, the sufficiency of this information to answer each of the questions that underpin those IPs varies from case to case and from IP to IP. The data suggest that some of this variation is at least part accounted for by the presence or absence of early triage. This is because the use of early triage is associated with enhanced operational performance in the form of increased ability of digital forensic methods to address all the seven Impact Points (see Figures 3.4 and 3.6 and Table 3.1).

It is also valuable to note that there is likewise an association between the use of early triage and enhancements in *Timeliness*. This is expressed by the decrease in the typical estimated time taken

[38] The chi-square test of association with Yates' continuity correction gives $\chi^2 = 24.6, p = 6.90 \times 10^{-7}$.

to answer the Questions Posed shown in Table 3.2 that is concomitant with the presence of early triage.

The performance enhancements noted in the previous two paragraphs are well explained by the ability of early triage to find evidence of criminality and to do so during those crucial early stages of an investigation. If it were known with certainty that this explanation is operative it would mean that there is a causative link between the use of early triage and those effects. However, as explored in Section 4.1.2, the observational nature of the Safenet study means that the possibility of a common cause explanation cannot be discounted.

Please also see Section 4.1.1 for a discussion of the utility of the Safenet data in the understanding of forensic value.

3.2 Place of early triage and the time between scene attendance and early interview

This work was carried out to address the research question: 'Does place of early triage have an effect on the time between scene attendance and early interview?'. It used the subset of the Safenet dataset that is described in Section 2.2.2. It produced the descriptive statistics shown in Tables 3.4 and 3.5 and the box and whisker plots[39] shown in Figures 3.10 and 3.11.

Table 3.4 Descriptive statistics for the time between scene attendance and early interview, grouped by the place of early triage (i.e., Scene, Station and HQ), with time expressed in hours.

Scene

n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
71	0.5	2.88	3.8	194.6	5.25	7564.52

Station

n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
20	1.53	3.7	4.34	454.11	6.34	8757.25

HQ

n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
16	1.07	3.89	5.78	396.57	15.6	4417.58

Table 3.5 Descriptive statistics for the time (expressed in hours) between scene attendance and early interview, grouped according to whether early triage occurred at the scene on the one hand or elsewhere (i.e., Station or HQ) on the other.

Scene

n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
71	0.5	2.88	3.8	194.6	5.25	7564.52

Elsewhere

n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
36	1.07	3.7	4.89	428.54	7.26	8757.25

[39] For information on box and whisker plots, and their interpretation, see Galarnyk (2018)

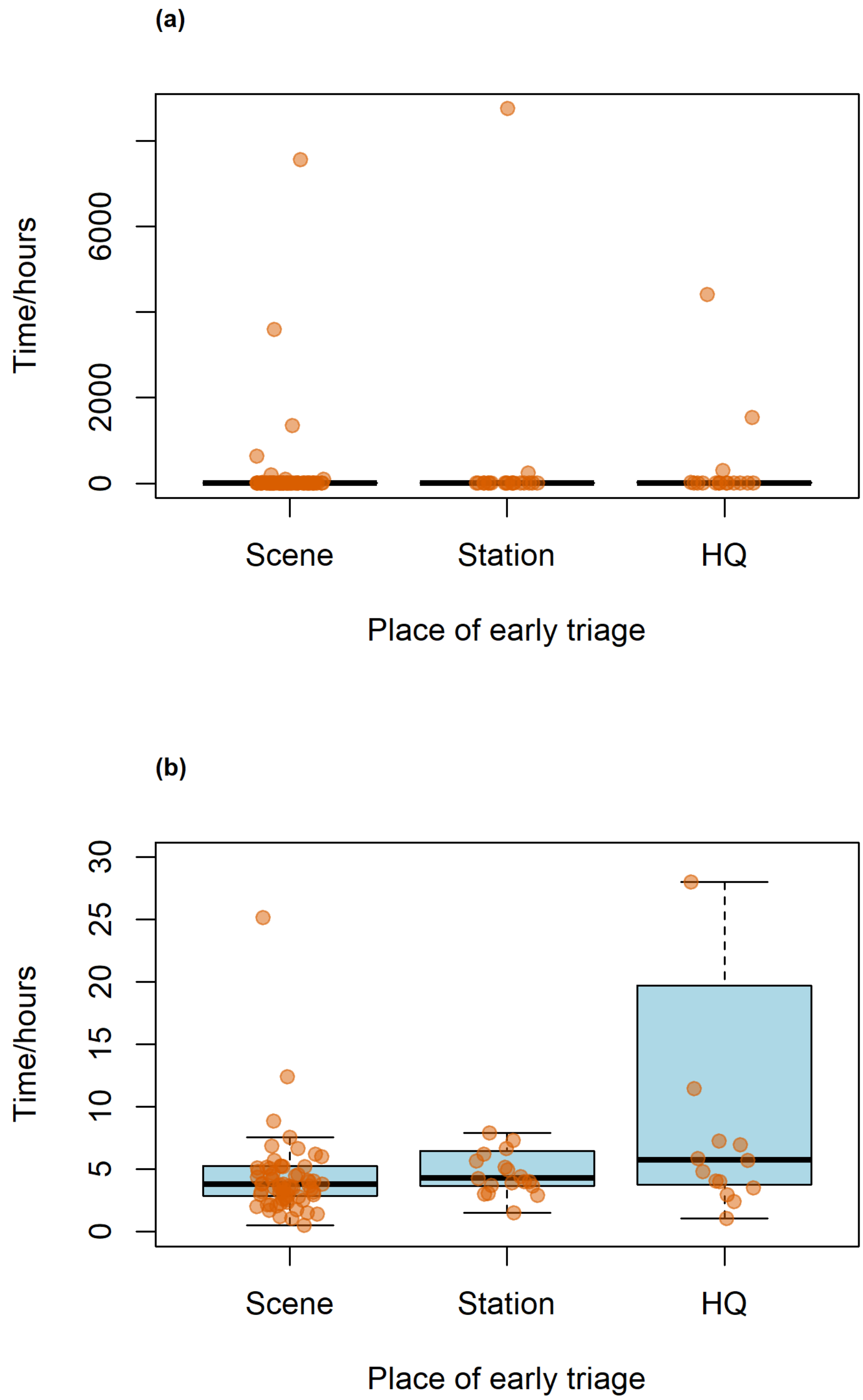
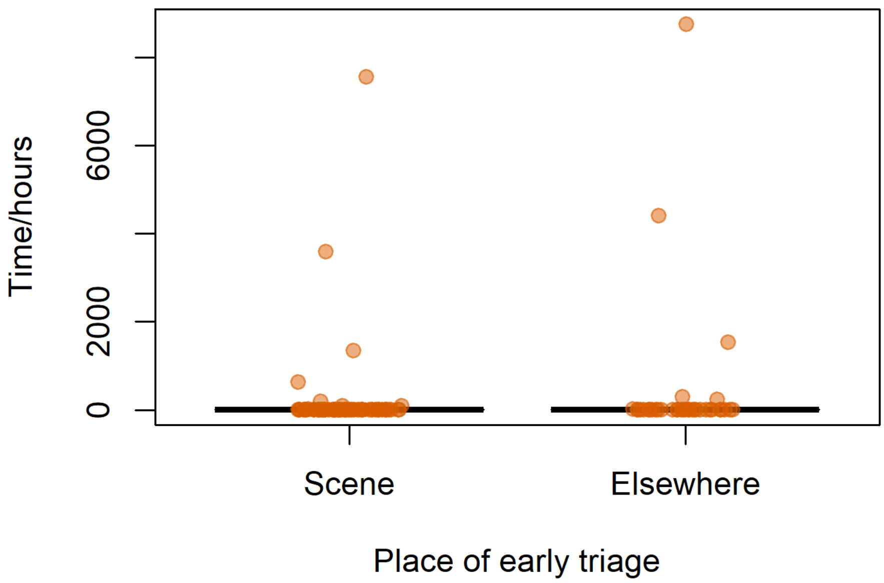


Figure 3.10 Box and whisker plots, with the raw data shown as dots, of the time between scene attendance and early interview, grouped by place of early triage. Both (a) and (b) are based on the full set of prepared data. However, the *y*-axis in (b) is limited to 30 hours to exclude the extreme outliers and to allow the full features of the box and whisker plots to be seen. Within each group (i.e., Scene, Station and HQ), the data points have been spread out on the *x*-axis merely to decrease the amount of overlap between those points.

(a)



(b)

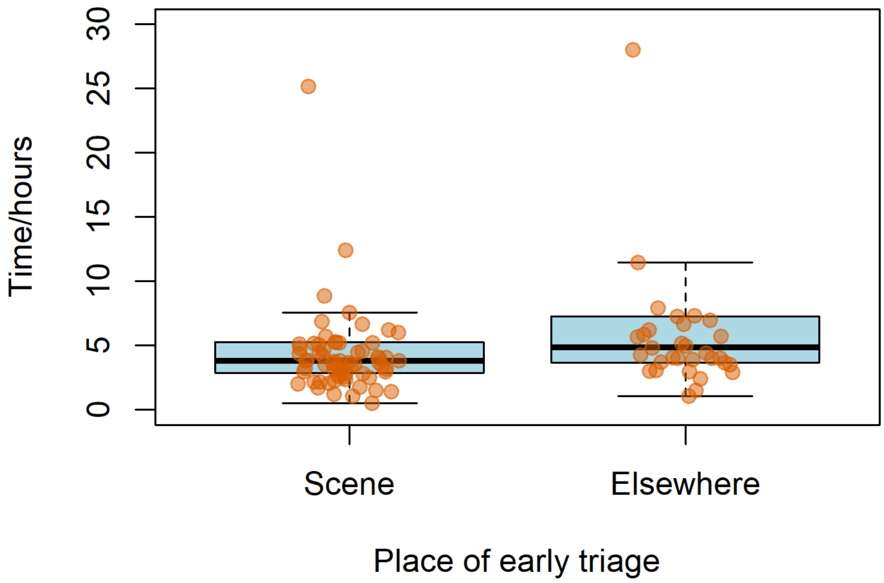


Figure 3.11 Box and whisker plots, with the raw data shown as dots, of the time between scene attendance and early interview, grouped by place of early triage. Both (a) and (b) are based on the full set of prepared data. However, the *y*-axis in (b) is limited to 30 hours to exclude the extreme outliers and to allow the full features of the box and whisker plots to be seen. In this case, Scene is as in Figure 3.10 and Elsewhere is an aggregate of the data labelled Station and HQ in that Figure. Within each group (i.e., Scene and Elsewhere), the data points have been spread out on the x-axis merely to decrease the amount of overlap between those points.

From the data shown in Tables 3.4 and 3.5 and in Figures 3.10 and 3.11, an association can be seen between the place of early triage and the typical (median) time between scene attendance and early interview; with this median time being shortest when early triage takes place at the scene. When the Station and HQ data are aggregated into one group (i.e., that labelled Elsewhere), that median is 1.09 hours, or 66 minutes, shorter when early triage takes place at the scene. This is a saving of 22.3%.

There are three possible explanations of the effect described in the previous paragraph, namely that:

- it happened by chance alone;
- whatever causes the place of triage to change from one police scene attendance to another also causes the change in median time between that attendance and early interview. This is the so-called common cause explanation;
- changing the place of triage causes a change in that median time. This is the so-called causal effect.

It would be very useful to tell these apart. This is because that knowledge allows the impact of change in early triage location on the typical time between scene attendance and early interview to be understood. Consider a scenario in which early triage is always carried out and in which a change in policy to one that mandates that early triage must be conducted at the scene is about to be implemented. If the last explanation is true and all else is equal either side of that policy change, its implementation can be confidently expected to lead to a shortening of that time. However, if either of the other two explanations accurately describes what is happening, any such policy change would not be expected to lead to that improvement in timeliness.

Statistical testing has been carried out. It can be argued that this provides some evidence to suggest that chance alone may not be a plausible explanation for the effect seen and yet, that evidence falls short of what is normally considered acceptable[40]. This means that further work would have to be done before the possibility of chance as the sole cause of the association seen could reasonably be ruled out. Even if this could be ruled out, more work would still be needed to distinguish between the remaining two explanations. It would be possible to devise an experimental study that could address these matters.

In the absence of the above-mentioned further work, it cannot be known what causes the effect of the place of early triage that is observed in this part of the Safenet study. Nonetheless, by the application of the precautionary principle, it is possible to use that observation make the recommendation given in the next paragraph.

From the above, it is recommended that where early triage is used, if the timeliness of the early interview is of importance, such triage should be conducted at the scene, not elsewhere. This is on the proviso that this option is available and that any risks associated with conducting early triage at the scene do not outweigh the possible benefits of doing so.

---

[40] Application of the Kruskal-Wallis test to the effect shown in Figure 3.10 reveals χ^2 = 5.28 and p = 0.071. Application of the Mann-Whitney U test to the effect shown in Figure 3.11 reveals W = 946 and p = 0.029. The second of these p-values is less than 0.05, making the effect seen significant at a threshold confidence level of 95%. This could be taken as evidence that that effect did not happen by chance alone. However, doing so would not take into account the familywise error rate inflation that occurs when the same data is repeatedly tested to address the same research question. Application of Bonferroni adjustment to that p-value to account for that inflation doubles it from 0.029 to 0.058, which is greater than 0.05. This means that it is no longer statistically significant at a 95% confidence level and so fails to reach the normally accepted threshold for the rejection of the explanation that the observed effect happened by chance alone.

3.3 Early triage and the early admission of guilt

This sub-study was carried out to address the research question: 'Does early triage lead to early admission of guilt?'

3.3.1 The overall effect of early triage

The contingency table shown in Table 3.6 was created by the cross-categorisation of the variables 'Early triage' and 'Early interview', which are as defined in Section 2.2.3. Table 3.7 shows the cell values of Table 3.6, each expressed as percentages of its row sum and Figure 3.12 shows those percentages in a graphical form.

Table 3.6 Counts of 'Early triage' (rows), grouped by 'Early interview' (columns)

	Admission	Non-admission	No interview	Sum
Positive triage	55	52	1	108
Negative triage	15	20	15	50
No triage	20	18	11	49
Sum	90	90	27	207

Table 3.7 Cell values of Table 3.6, each expressed as percentages of its row sum (individual percentages are rounded to the nearest 0.1 percentage point)

	Admission	Non-admission	No interview	Sum
Positive triage	50.9	48.1	0.9	100
Negative triage	30.0	40.0	30.0	100
No triage	40.8	36.7	22.4	100

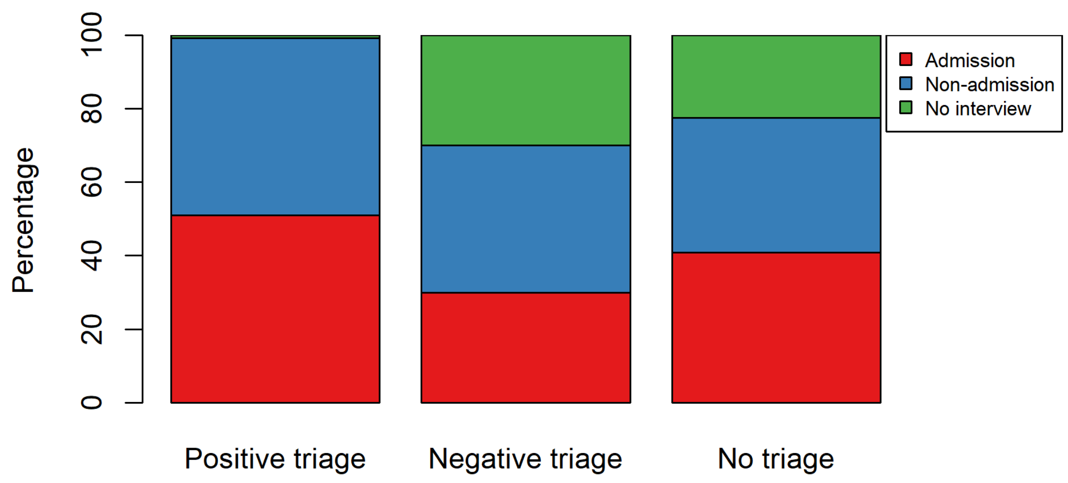


Figure 3.12 Cell values of Table 3.6, each expressed as percentages of its row sum

Chi-square testing of the data shown in Table 3.6 shows an effect that is highly statistically significant ($\chi^2 = 31.24$, $p = 2.7 \times 10^{-6}$), suggesting that it did not happen by chance. However, this detection of a significant effect does not provide information concerning which of the nine cells of Table 3.6 are responsible for it. To find this information, follow-up (aka *post hoc*) testing was carried out to calculate the Bonferroni corrected *p*-value associated with each of those cells.

It is evident from Figure 3.12 that, in the sample, the overall proportion of opportunities that produced an early admission of guilt is noticeably larger amongst those with a positive early triage result than for those with either a negative early triage finding or in which early triage did not occur. In the sample, the percentage of *early interviews* that result in an early admission of guilt (as defined in Section 2.2.3) for 'Positive triage', 'Negative triage' and 'No triage' are calculated to be respectively 51.4%, 42.9% and 52.6%. For the reasons set out in Section 3.3.3, this does *not* supply sufficient evidence to suggest that the reason why a positive early triage finding is associated with an increased frequency of early admissions of guilt is because such a finding leads to a greater likelihood of the admission of guilt during an early interview.

Importantly, from Figure 3.12, it can also be observed that, in the sample, 'Positive triage' and 'Negative triage' are respectively associated with low and high incidences of 'No interview'. These observations are key because early admission of guilt can only happen in an early interview — if there is no early interview, there can be no early admission.

Crucially, the follow-up testing shows that these key observations are the only ones in the nine cells of Table 3.6 that are statistically significant. For 'Positive triage' the relevant *p*-values are 5.8×10^{-7} when corrected for the follow-up testing referred to above and 8.4×10^{-7} when corrected for all relevant tests that were carried out. For 'Negative triage', the corresponding figures are 3.9×10^{-4} and 5.7×10^{-4}. All four of these figures are << 0.05, indicating a statistically highly significant effect.

3.3.2 The effect of early triage occurrence on early admission of guilt in those opportunities with an early interview.

The contingency table shown in Table 3.8 was created by the cross-categorisation of the variables 'Early triage occurrence' and 'Early admission', which are as defined in Section 2.2.3. Table 3.9 shows the cell values of Table 3.8, each expressed as percentages of its row sum and Figure 3.13 shows those percentages in a graphical form.

Table 3.8 Counts of 'Early triage occurrence' (rows), grouped by 'Early admission' (columns)

	Admission	Non-admission	Sum
Triage occurred	70	72	142
No triage	20	18	38
Sum	90	90	180

Table 3.9 Cell values of Table 3.8, each expressed as percentages of its row sum (individual percentages are rounded to the nearest 0.1 percentage point)

	Admission	Non-admission	Sum
Triage occurred	49.3	50.7	100
No triage	52.6	47.4	100

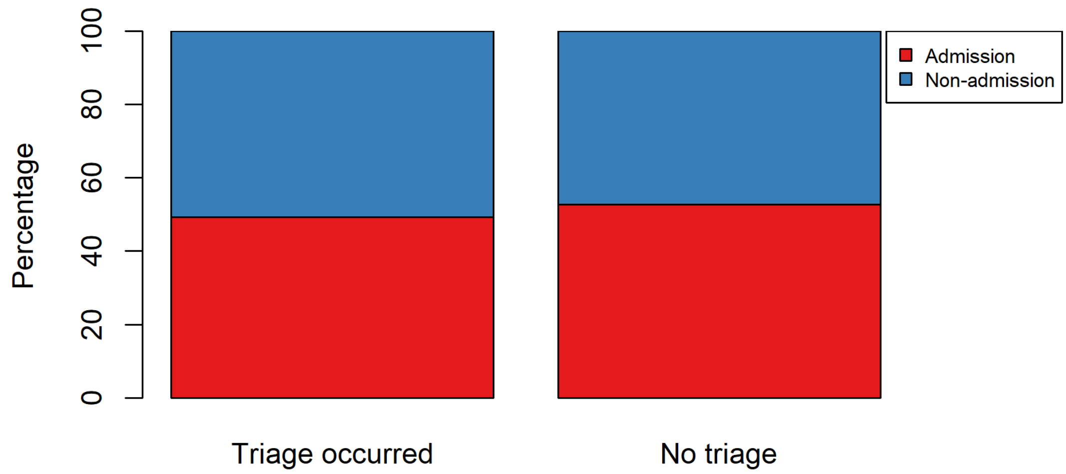


Figure 3.13 Cell values of Table 3.8, each expressed as percentages of its row sum

In the sample of 207 opportunities that are analysed in Section 3.3.1, there are 27 in which no early interview occurred. This leaves 180 opportunities to be analysed in this Section (i.e., Section 3.3.2).

From the metrics given in Table 3.9, it can be seen that in that sample of 180 opportunities:
- an early admission of guilt occurred in 52.6% of those opportunities where early triage did not happen;
- this figure fell to 49.3% amongst the opportunities where early triage did occur.

This effect can reasonably be considered to be very small[41]. Pearson's chi-square testing reveals that it is *not* statistically significant ($\chi^2 = 0.03$, uncorrected-p = 0.86, Bonferroni-corrected-p = 1), suggesting that it happened by chance.

From this, it can be concluded that there is no evidence to suggest that the introduction of early triage leads to admission of guilt during early interview.

3.3.3 The effect of early triage finding on early admission of guilt

The contingency table shown in Table 3.10 was created by the cross-categorisation of the variables 'Early triage finding' and 'Early admission', which are as defined in Section 2.2.3. Table 3.11 shows the cell values of Table 3.10, each expressed as percentages of its row sum and Figure 3.14 shows those percentages in a graphical form.

Table 3.10 Counts of 'Early triage finding' (rows), grouped by 'Early admission' (columns)

	Admission	Non-admission	Sum
Positive	55	52	107
Negative	15	20	35
Sum	70	72	142

[41] The **w**-value for this effect is 0.027 and Cohen (1988) suggests that the values **w** = 0.1, 0.3 and 0.5 can be thought of as being small, medium and large effect sizes, respectively.

Table 3.11 Cell values of Table 3.10, each expressed as percentages of its row sum (individual percentages are rounded to the nearest 0.1 percentage point)

	Admission	Non-admission	Sum
Positive	51.4	48.6	100
Negative	42.9	57.1	100

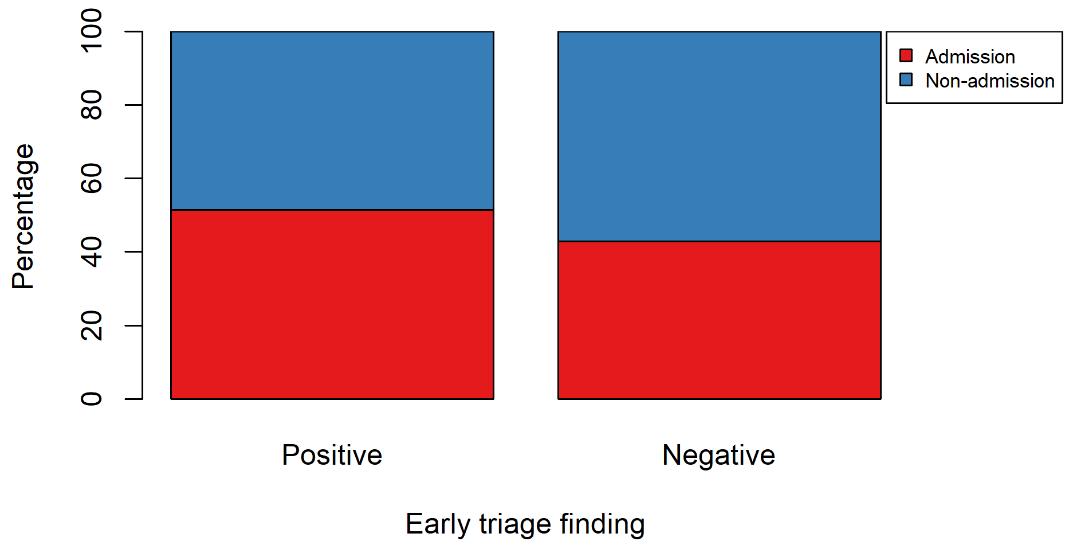


Figure 3.14 Cell values of Table 3.10, each expressed as percentages of its row sum

In the sample of 207 opportunities that are analysed in Section 3.3.1, there are 65 for which there was no early triage or no early interview or both. This leaves 142 opportunities to be analysed in this Section (i.e., Section 3.3.3).

From the metrics given in Table 3.11, it can be seen that in that sample of 142 opportunities:

- an early admission of guilt occurred in 42.9% of those opportunities where early triage finding was negative;
- this figure rose to 51.4% amongst the opportunities where early triage finding was positive.

This effect can reasonably be considered to be small[42]. Pearson's chi-square testing reveals that it is *not* statistically significant ($\chi^2 = 0.47$, uncorrected-p = 0.49, Bonferroni-corrected-p = 1), this suggests that the effect seen in the sample may have happened by chance alone.

From this, it can be concluded that there is no evidence strong enough to suggest that a positive early triage finding is more likely to be followed by an admission of guilt at early interview than is a negative early triage finding. However, it is interesting to note that:

1. Moston and Engelberg (2011) state '*The suspect's decision to confess during questioning by police may be influenced by a range of factors*', and, '*by far the strongest single predictor of a confession is the strength of evidence against the suspect*

[42] The **w**-value for this effect is 0.074 and Cohen (1988) suggests that the values **w** = 0.1, 0.3 and 0.5 can be thought of as being small, medium and large effect sizes, respectively.

…, a fact that has been demonstrated in studies employing a diverse range of methodologies.'.

2. As stated in Section 2.2.3, the definitions of positive and negative triage findings only concern whether, in any given case, evidence of criminality was found on one or more mobile phones, tablets or SIM cards. This means that any opportunities in which early triage found evidence of criminal activity solely on digital devices that were not mobile phones, tablets or SIM cards would be categorised as producing a negative triage finding. It is not known how often this happened as this data was not available. If it did occur, it could mask any increased likelihood of the admission of guilt at early interview that might accompany the finding of evidence of criminality during early triage.

3.3.4 Conclusion

As can be seen from Section 3.3.2, that there is no evidence to suggest that the introduction of early triage leads to admission of guilt during early interview.

As described in Section 3.3.3, it is true that, amongst the data analysed, the frequency of admissions of guilt during early interview is higher when there was a positive early triage finding rather than a negative one. However, this effect is not statistically significant, meaning that it is plausible that it happened by chance alone. But, it must be noted that, as discussed in Section 3.3.3, there are circumstances that could exist in which the discovery of evidence of criminality during early triage could be categorised as producing a negative early triage finding. If such circumstances occurred amongst the opportunities that were analysed, this would downplay the effect referred to earlier in this paragraph.

As described in Section 3.3.1, positive early triage, as opposed to negative early triage, *is* associated with an increased likelihood of early interview occurrence. This is important because admission of guilt can only happen during an interview[43].

3.4 The time between referral and scene attendance, and the loss of digital evidence

This sub-study was carried out to address the research question: 'Does delay between the date of referral and the date of scene attendance lead to loss of digital evidence?'. In this context, referral means the police's receipt of the intelligence report that alerted them to the possibility of a case involving online child sexual exploitation (CSE).

To answer this research question with certainty requires knowledge of cases where digital evidence was present at the time of referral but which was then lost at a known point in time before scene attendance. This information is not available and so inference in the face of uncertainty must be used if the question is to be addressed.

The data analysed consists of 4 samples, each of which are subsets of the complete Safenet dataset (please see Section 2.2.4 for details). Each of these samples contains between 156 and 215 rows, each of which records an instance (opportunity) when scene attendance provided the possibility of the discovery of digital evidence.

[43] It is perhaps interesting that there is a *hint* that, had more data been collected, it could have detected that the use of early triage is associated with an increased likelihood of an early interview. For details, please see Section 6.2.1 of Appendix 8.

There are 4 samples because, as detailed in Section 2.2.4, this sub-study uses 4 different indicators of digital evidence discovery. Each of these indicators is a variable with possible values of 'Yes' and 'No'. For each indicator, the value 'Yes' was assigned to each opportunity in the relevant sample where the indicator detects that digital evidence was discovered and 'No' to each where it does not. The value 'No' is an indication of the *possible* loss of digital evidence to the criminal justice system (CJS).

For each of the opportunities in each of the samples, the time interval between the dates of referral and scene attendance was also established. In this sub-study, the inferential approach is to see whether the sample data suggests that, at the population level, the value of any given indicator tends to alter with changes in that time interval.

Imagine a scenario in which:

- such tendency is suggested and
- the median value of that time interval is smaller amongst those opportunities in the sample where the value of the indicator is 'Yes'.

Under these circumstances, the suggestion is that the answer to the research question is 'Yes'.

However, this suggestion will only be dependable if:

- the indicator used reliably detects the loss of digital evidence to the CJS and
- there is no mechanism at play that causes both increases in the time interval and that loss.

To illustrate why the second of these provisos is necessary, consider what could happen if low priority were given to cases where it is judged unlikely that digital evidence will be found. If this lower priority delays the date of scene attendance and if that judgement turns out to be correct, this would lead to a negative association between the time interval and the likelihood of the discovery of digital evidence. The correct understanding of this association would be that decreases in the probability of digital evidence discovery leads to longer time differences between referral and scene attendance, not the other way around.

We now know that there are circumstances in which it would be possible to suggest that the answer to the research question is 'Yes'.

Now consider what might be inferred if it is found that there is insufficient reason to suggest that, at the population level, the median time between referral and scene attendance is smaller amongst opportunities where indicator value is 'Yes'. This does not necessarily suggest that the answer to the research question is 'No'. This is because there could be influences at play, other than the time between referral and scene attendance, which alter the likelihood of a given indicator value being produced and these influences hide the effect of that time interval.

To illustrate this, by way of an example and for argument's sake, let us assume that the likelihood of finding no digital evidence is greater when a crime has not been committed than when it has. Consider what effect this could have if, due to chance or some other reason, this type of occurrence is not evenly distributed amongst the time intervals between referral and scene attendance that are seen in the sample. Under these circumstances, whether a crime has been committed could provide an influence of the type referred to in the last sentence of the previous paragraph. To avoid this possibility, as listed in Table 2.5 (in Section 2.2.4), all opportunities where the investigation found that the suspect had not committed any of the crime(s) they were suspected of committing were removed from the data prior to analysis.

With all of the above in mind, let us turn to the results of this sub-study, which are shown in Figures 3.15 to 3.18. From these it can be seen that, amongst the sample data, for all four indicators there is an association between the time interval from referral to scene attendance and the value of the indicator concerned. This association is summarised by the fact that the median value of that time interval[44] is smaller amongst those opportunities in the sample where the value of the indicator is 'Yes'.

For argument's sake, let us assume that that association is a reflection of a similar effect in the population and that each indicator reliably detects the loss of digital evidence. Under this scenario, at the population level, increased time between the date of referral and the date of scene attendance would be associated with an increased likelihood of the loss of digital evidence. Therefore, if the provisos stated earlier in this Section are in place, any such association would allow it to be inferred that the answer to the research question would be 'Yes'.

However, the association seen in each of Figures 3.15 to 3.18, inclusive is what might reasonably considered to be very small[45]. More importantly, none of those associations are statistically significant[46], indicating that a plausible explanation is that they have occurred by chance alone and that there is no such effect in the population. Also, it should be noted that it would not be reasonable to treat the observations presented in any one of those Figures as being independent of the others. This means that they should not be seen as confirming each other's findings. Instead, they are probably better seen as echoes or manifestations of the same underlying pattern in the data analysed.

[44] For each box and whisker plot, the median time interval of the relevant data is shown by the location of the thick black vertical line.

[45] The absolute value of r for the association seen in each of Figures 3.15 to 3.18 is, respectively, 0.07, 0.07, 0.09 and 0.05. Cohen (1988) provides guidelines that suggest that, for behavioural studies at least, r values of 0.1, 0.3 and 0.5 represent small, medium and large effect sizes, respectively.

[46] Using the data shown in each of Figures 3.15 to 3.18, inclusive, the Wilcoxon rank sum test with continuity correction reveals *p*-values of 0.41, 0.33, 0.19 and 0.43, respectively. Applying Bonferroni adjustment to allow for familywise inflation in the error rate produces corrected p-values of 1, 1, 0.77 and 1, respectively. Clearly, none of these *p*-values are < 0.05.

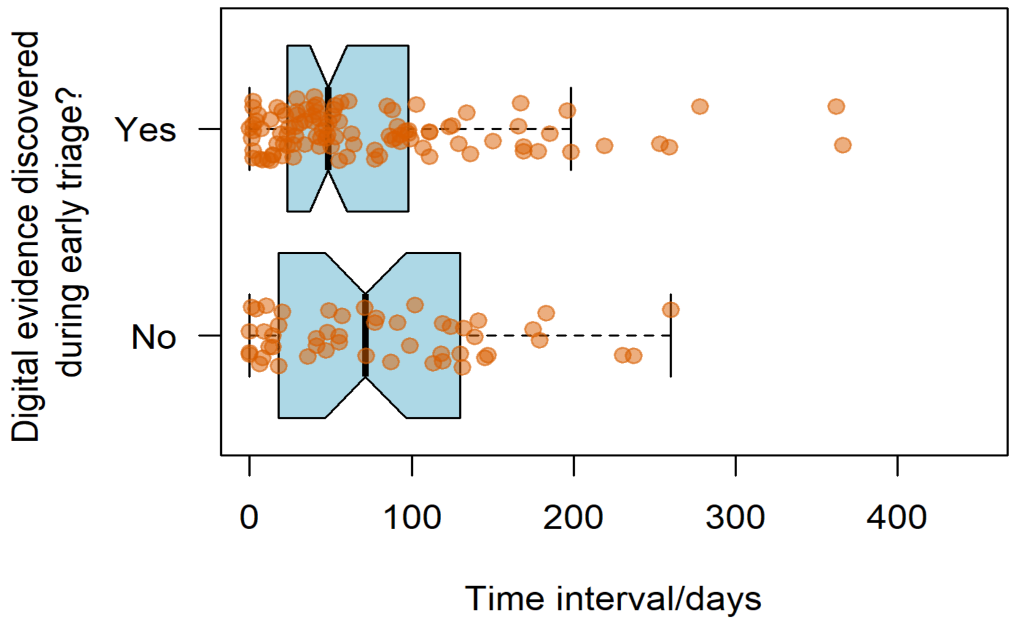


Figure 3.15 A plot of the time interval between the date of referral and the date of scene attendance, grouped by whether digital evidence was discovered during early triage.

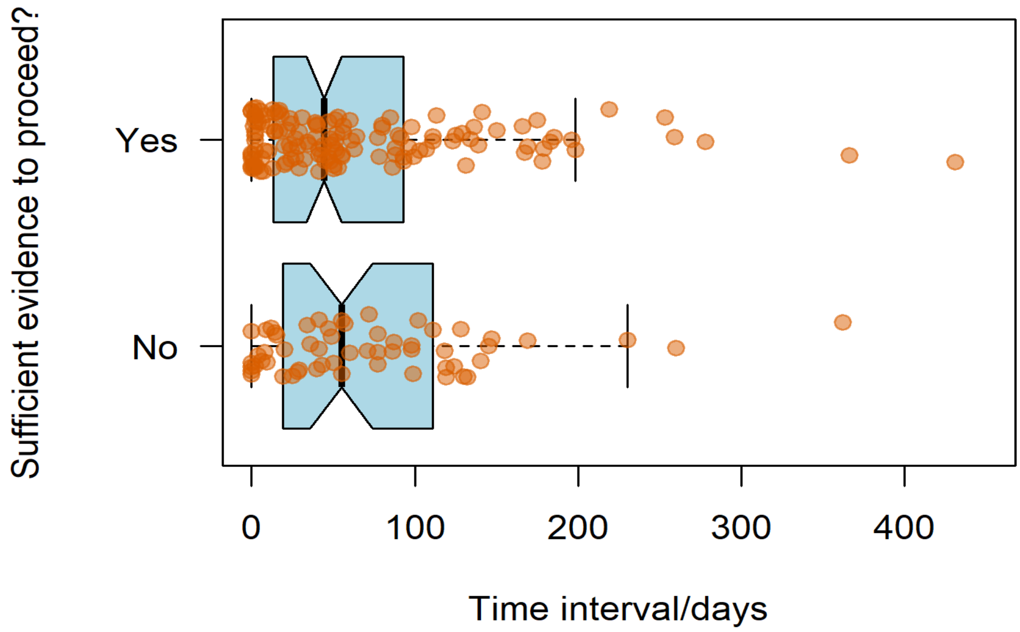


Figure 3.16 A plot of the time interval between the date of referral and the date of scene attendance, grouped by whether it was deemed that there was sufficient evidence to proceed.

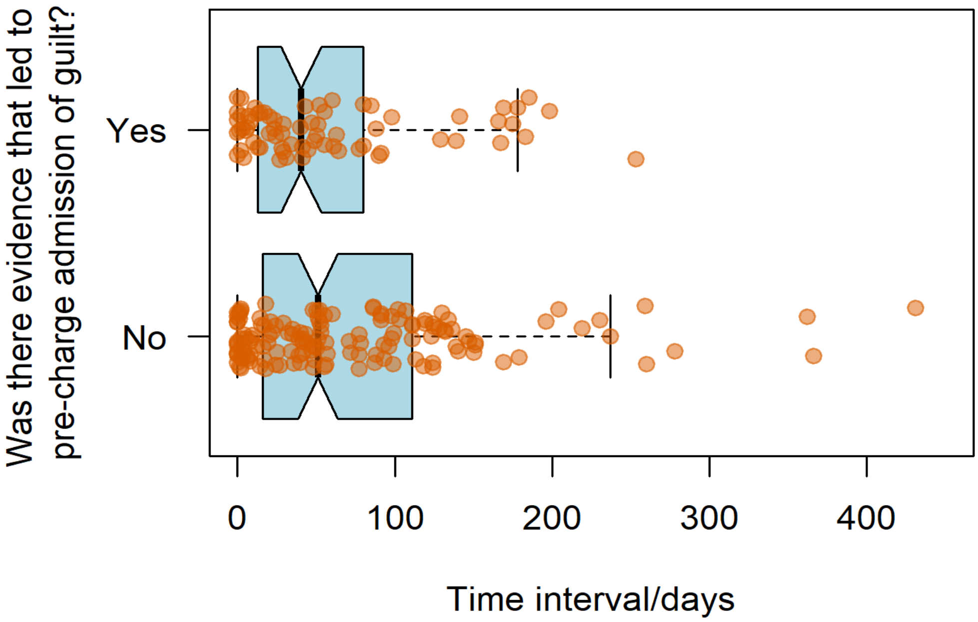


Figure 3.17 A plot of the time interval between the date of referral and the date of scene attendance, grouped by whether it was deemed that there was evidence that led to pre-charge admission of guilt.

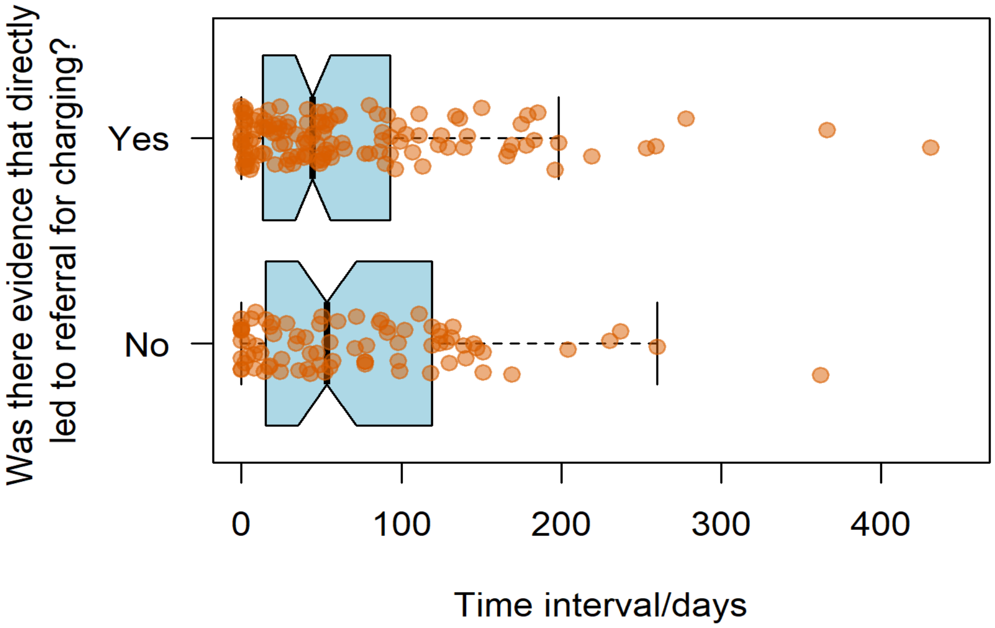


Figure 3.18 A plot of the time interval between the date of referral and the date of scene attendance, grouped by whether there was evidence that directly led to referral for charging.

Therefore, taking all of the arguments and data presented in this Section into account, it is not possible to give a definitive answer to this sub-study's research question. Nonetheless, even in the absence of such an answer, use of the precautionary principle allows the following conclusions to be drawn:

1. In intelligence initiated online CSE cases, undue delay between referral and scene attendance should be avoided.
2. However, if such delay occurs, the resourcing of the examination of devices for digital evidence should not be reduced.

Also, this sub-study has provided no evidence to suggest that the age of a digital device should be used as a criterion to decide whether to include it for examination. Finally, continued monitoring of any association between the time from referral and scene attendance and the loss of digital evidence is warranted. This is because changes in technology and/or the behaviour of criminals could alter the state of affairs that are revealed by this sub-study.

4. Evaluation and recommendations

4.1 The data's utility

4.1.1 The utility of the data in the understanding of forensic value

It has been recognised for some time that there is a clear need to better understand the contribution and value of forensic science to the UK's Criminal Justice System (Home Office 2016b, Home Office, APCC and NPCC 2019). The Home Office's Impact of Forensic Science project, of which this study is part, is a direct response to that need (Barrett 2021a and 2021b). That project has devised the Impact Point approach that is described in Section 1.5. This study has applied that approach to the measurement of the *Effectiveness* and *Timeliness* (as defined in that Section) of digital forensic methods. This has been done in the context of the combatting of online child sexual exploitation as practised by Staffordshire Police in its Operation Safenet (for details of that Operation, see Section 1.4). As shown in Section 3, this approach has successfully illuminated the contribution and value of those methods in that context. This provides a proof of principle for the Impact Point approach and demonstrates the clear utility of the data gathered in the understanding of forensic value. It has also illuminated the ability of the Impact Point approach to understand that value in unprecedented detail.

The useful structure of the data, which was designed by Barrett (2021b), is worthy of comment also. A number of the cells of the Safenet spreadsheet contain comments that provide additional data. These aside, the Safenet dataset is compliant with Wickham's concept of tidy data (Wickham 2014). A key feature of this concept, and of the Safenet dataset except those comments, is the rigorous use of rows and columns to respectively represent variables and observations. This is highly advantageous. This is because it makes the data readily accessible to analysis via software such as Excel, as used by Barrett and Varraich in the Impact of Forensic Science project as a whole (Barrett and Varraich 2021), and the statistical programming language R, as used in the Safenet study which is reported here. Also, as listed in Appendix 3, drop-down menus have been used to record the values of many of the Safenet dataset's the categorical variables. This eliminates typographical errors during the entry of such data, which significantly decreases the amount of data processing prior to analysis.

The richness of the Safenet data means that much of it remains unanalysed. Such analysis, coupled as needs be with the collection of additional data, could be used to allow the Impact Point approach to produce a yet deeper understanding of forensic value. Examples of research questions, all of which are specifically concerned with Operation Safenet, that could be accessible by such means include:

- 'How do the costs and benefits of forensic science compare?' (The *Cost Benefit* Impact Metric is already part of the Impact Point approach (see Section 1.5) but has not been explored in the study reported here.)
- 'What is the *Effectiveness*, *Timeliness* and *Cost Benefit* of the use of digital forensic methods for all 27 of the Impact Points (IPs) that have been identified by the Impact of Forensic Science project?' (As detailed in Table 1.1, the Safenet study is only concerned with a subset of those IPs. For a complete list of all 27, along with their definitions, see Appendix 1, which is from Barrett [2021b].)
- 'Can the circumstances under which digital forensic methods provide all of the information needed to address a given IP be understood?'
- 'Can the reasons for the variation in the utility of digital forensic methods from one IP to the next that is evident in Figure 3.2 be understood?'

The proof-of-principle study reported here is, in effect, a pilot. It has allowed the largely untapped potential of the Impact Point approach to be recognised. However, it has highlighted areas for development and lessons to be learnt. These are explored in Sections 4.1.2, 4.2 and 4.3.

4.1.2 The utility of the data in the understanding of cause and effect

This Section provides an analysis of how useful the Safenet data is for the understanding of cause and effect, and how that utility can be improved. For the interested reader who has little or no familiarity with how such understanding can be achieved, it also provides the theoretical underpinning of that analysis. As a consequence, it is rather didactic. A succinct but sufficiently comprehensive summary of that analysis is provided in Section B.2 of the Executive Summary.

Firstly, let us define some terms that will help us discuss this matter. Two variables are associated if changes in the values of one are mirrored in some way by variations in the values of the other. When an association is present an effect has occurred. All possible data of interest that could exist is called the population. The data used in a study is a subset of the population and is called the sample.

There are two broad categories of study that can be used to detect associations between two variables of interest. These are experimental studies, such as Randomised Control Trials (RCTs), and observational studies. In both cases, the value of one of the variables is monitored as the value of the other variable changes. The difference is that in experimental studies, the researcher intervenes to assign specific values of that other variable to specific instances, whereas, in observational studies, advantage is taken of variation in the value of that variable amongst instances as they happen. In any one study of either type, those instances will all be of a similar ilk and could be, for example people, fields of crops, trees, cases of a particular disease etc. In the Safenet study, each instance is an opportunity that digital forensics had to attain one of the seven Impact Points (IPs) by addressing the corresponding Question Posed (QP) (for details of the IPs and QPs, see Section 1.5). In either type of study, an association is found in the sample if variations in the value of one variable of interest alter in a systematic way with the value of the other.

Whenever an association is seen in a sample, there are three possible explanations. One of these is that it happened by chance alone. In other words, the associations seen in the sample is not a reflection of an association of the same type present in the population. Instead, it has been formed by the action of chance during the subsetting process that created the sample from the population.

It is important to be able to detect such occurrences. This is because associations seen in the sample that are due to chance alone cannot be expected to be reproducible. This means that any policy change that was based on the observation of such a chance association will fail to produce its intended benefits.

If the effect of the association seen in the sample is large and the number of data points (i.e., the sample size) is substantial, this tells against the plausibility that the effect seen in the sample has occurred by chance alone. Statistical testing of the type described to in Section 2.2.5 can be used to help quantify the confidence with which the possibility of chance as the sole explanation of the effect seen can be rejected. These tests are equally applicable to both experimental and observational studies and have been used in the study reported here.

If it is deemed that the effect seen in the sample has not happened by chance alone, two possibilities remain. These are:

1. that whatever caused the change in one of the variables also influenced the value of the other. This is known as the common cause explanation.
2. There is a causal association between the variables. Statistically speaking, a causal association is one where intervening to alter one of the variables results in a systematic change in the likely value of the other.

It is important to tell these apart. To illustrate this, consider a scenario in which fluctuations in the workload of the Digital Forensics Unit (DFU) influence both whether early triage is undertaken and the amount of time it takes to process items seized from scenes. Under these circumstances that influence would act as a common cause that could lead to an association between the use of early triage and improved *Timeliness* being seen in the sample. It is entirely possible that statistical testing would lead to the conclusion that this was not a chance occurrence. If, on this basis, policy were changed to make early triage mandatory in the expectation that this would lead to improved *Timeliness* it can be anticipated that that expectation would not happen. This is because it is changes in workload that caused the effect seen, not changes in whether early triage was used.

RCTs are designed to provide an environment where causative associations occur but those with a common cause do not. Crucially, such trials each involve more than one group of instances where an effect could be seen. Importantly, in an RCT, the instances are randomly assigned to the groups. In the simplest form of RCT, there are two groups. The intervention of interest (e.g., the use of early triage) is applied to the members of one of these (the experimental group) whilst those in the other (the control) is left free of that intervention. Then, a suitable quantitative indicator of the effect of interest is monitored in both groups (e.g., whether a suspect admits guilt during an early interview). It is highly likely that some form of association between the presence or absence of the intervention and the degree of change detected by the indicator in the sample will be observed. If use of the previously mentioned statistical methods indicates that this association has not happened by chance alone, the presence of the random allocation process described above means that this outcome provides direct evidence of a causal association[47].

Policy change is carried out with the specific intention of using an intervention to bring about a desired causal effect. Prior to policy change roll out, well-designed RCTs can be used to provide evidence of whether any given intended intervention does bring about the hoped-for effect and, if so, give an estimate of its size.

It is for the reasons set out above that RCTs are often seen as the 'gold standard' for the detection of causal effects. This is not to say that RCTs are infallible, nor would it be appropriate to believe that observational studies cannot help establish causal associations; it is just that much greater caution needs to be taken when using data from observational studies to infer causation. Also, RCTs are not always appropriate. For example, they may be unethical or may take too long to complete. When they are not practicable, observational studies can be the best choice.

In the case of the Safenet study, the principal reasons for the use of an observational approach are:

- had an experimental design been used, the data would have had to be gathered during the disruptive effects of the COVID-19 pandemic, meaning that any lessons learnt might not be applicable at other times;
- by using an observational study, data spanning multiple years could be collected. This allowed the inclusion of *many* more cases with complete paths through the Criminal Justice System than would have been feasible if an experimental approach had been taken.

---

[47] For an excellent and very readable introduction to statistical approaches to the detection of cause and effect (amongst other relevant topics) see Spiegelhalter (2019).

Also, the Safenet study is part of a wider proof of principle study designed to see whether the Impact Point approach, set out in Section 1.5, can be used to gauge the value of forensic science. So, its primary purpose was not to detect causal effects.

There was, however, variation in practice that is captured in that dataset. Most notably and as previously mentioned, there are instances where early triage was used and others where it was not. This means that instances with and without early triage can be seen as forming separate groups. It is tempting to see these as substitutes for the experimental and control groups of an RCT designed to test whether the introduction of early triage has effects of interest. However, such substitution has to be done with great caution because a key feature of RCTs is missing. That feature is the random allocation of instances to the control and experimental groups. This leaves the possibility of a common cause explanation for any associations that were observed.

Section 3 contains clear evidence of associations between the use of early triage and multiple operational enhancements. As discussed there, these enhancements are well explained by early triage's ability to provide evidence of criminality during the crucial early hours of an investigation. This provides mechanisms by which early triage could cause these valuable effects. However, for the reasons explained above, the observational nature of the Safenet study means that the existence of common cause explanation(s) cannot be ruled out.

For operational reasons, it would be valuable to know with greater surety the likely primary driving force(s) behind the beneficial effects that are associated with early triage. There are several approaches that could throw light on this. Principal amongst these are the use of:

- statistical techniques, such as propensity score analysis (Lalani et al. 2020) and multivariable regression modelling, to attempt to deal with any bias caused by the existence of common cause effects;
- an RCT to test the apparent effects of early triage;
- qualitative research informed by business process analysis to gather evidence that could differentiate between the explanations set out in the previous paragraph.

The first of these would be the easiest to deploy if, serendipitously, the Safenet dataset already contains the necessary data. However, it is not guaranteed to succeed. The second would be the preferred option if the overriding consideration were to find whether there is a causal association between early triage and its apparent effects. Even if such a trail were to establish such a causative links, it would not necessarily unearth the mechanism(s) of that causation. Such mechanism(s) could be operationally important and should not be assumed. The third approach on the list (i.e., qualitative research) would be directly aimed at finding the mechanism(s) that explain the associations seen. It would also be relatively quick and straightforward to conduct and would not be affected by the perturbations caused by the current pandemic. Also, its use would not preclude the use of an RCT at a later date. For these reasons, it is currently the preferred option and work has already started on such a study.

Once a beneficial effect of a choice that is within operational control has been detected and its mechanism has been understood, there is a further piece of information that must be established before that knowledge can be used to optimise performance. This is the nature and extent of any unwanted side effects of making that choice. This is a topic that is returned to in Section 4.2.3.

4.2 Further development of the Safenet dataset concept

4.2.1. Data capture

Excel was used to capture the Safenet data. Excel's ease of use, its widespread availability and familiarity, its inbuilt data analysis functionality and the ease with which its files can be saved in csv format all make this a good choice for use in a pilot study such as this. For the same reasons, it would also be a good choice for any future "dip-tests" or relatively small-scale pilot studies that require manual data entry.

Excel also has a 'Data Validation' tool[48] that can be used to reduce the opportunity for data entry error. It does this by allowing restrictions to be placed on the data types or values that can be entered into any given cell of an Excel spreadsheet. For example, this can be achieved by the creation of drop-down menus, which limit what can be entered into the cell concerned to one of several possible allowed values. Such menus are very useful for categorical data, such as the presence or absence of Digital Forensic Unit staff during scene processing, and were employed to good effect in the Safenet dataset creation. However, they are not applicable to continuous data, such as dates and times. Neither are they useful for frequency data, such as counts of the number of exhibits seized, unless the range of expected values is both known and very restricted. Nor do they assist in eliminating errors that could be detected because of logical impossibilities, such as would occur if the number of exhibits analysed is recorded as exceeding the number seized or if the analysis of exhibits is recorded as preceding their seizure. Excel's Data Validation tool can also be used to forbid data entry that creates such illogicalities.

Larger-scale studies would benefit hugely from automated or semi-automated data capture and work is planned to explore this possibility further. The process of gathering the Safenet data has provided information that will be of value to this work.

Based on the lessons learnt from the data capture process used in the study reported here, the following recommendations are made:

- Establish a project-wide formal framework for data quality control;
- Ensure that the collected data are available in a file format, such as csv, that is readily accessed by the wide range of software used for data analysis (Excel, R, SPSS, Python, SAS etc).
- Ensure that each release of raw data has a time stamp and a unique filename. That file or raw data should be treated as immutable by the analysts. Any changes to the dataset made by them should be carried out on copies of the raw data.
- Consideration should be given to the project-wide adoption of a formal version control system such as git[49].
- Increase the number of Impact Points (IPs). Initially, this would be to include all 27 that have been developed by the Home Office's Impact of Forensic Science project (see Appendix 1 for details, which is from Barrett [2021b]). Further expansion of that list could be valuable. For example, IPs that specifically allow the measurement of victim experience and proportionality could be used to good effect.
- Refinement, as needs be, of the current Questions Posed and their precise definition.
- Create a standardised methodology for determining, in any given instance, whether any given investigative technique used, whether forensic or otherwise, has contributed to answering the Question Posed (QP). It would seem to be prudent to do this in collaboration

---

48 https://support.microsoft.com/en-us/office/more-on-data-validation-f38dee73-9900-4ca6-9301-8a5f6e1f0c4c (accessed 23 May 2021)

49 https://git-scm.com/ (accessed 29 August 2021)

with the Forensic Capability Network's expert network on performance and risk, which is made up of force forensic leads.

- Add variables to allow the:
    - ready extension of the *Timeliness* Impact Metric across all IPs used and improve the precision/accuracy of that metric;
    - calculation of the *Cost Benefit* Impact Metric.
- Minimise the opportunities for unconscious cognitive bias and, where possible, monitor and measure its impact.
- Remove redundant variables from the dataset.
- Widen data capture to multiple police forces. This will increase the speed with which sufficient data can be collected to allow effects to be detected and aid the anonymisation of data.
- Establish methods for the automated recording of data, with the necessary investment occurring at a national level.
- Facilitate the addition of further variables, and/or further levels to existing variables, to allow new research and policy development to occur.
- Widen the list of crime types and forensic methods used.
- Continue and extend the use of the `Tidy Data` (Wickham 2014) format. In the process, minimise the use of free text during data entry and discontinue the use of Excel comments to store data.

4.2.2 Further developments to aid operational management

As reported in Sections 2.2.1 and 3.1, Barrett and Varraich (2021) have produced a number of interactive dashboards from the Safenet data. Were these updated on a regular basis, such dashboards would allow ready access to synoptic statistics that could better inform operational management decisions.

The Safenet study was created as one of six proof-of-principle studies designed to test the utility of the Impact Point approach to the measurement of forensic value (Barrett 2021a and 2021b). To achieve that measurement, as described in Section 1.5, that approach proposes the use for three Impact Metrics, namely *Effectiveness*, *Timeliness* and *Cost Benefit*. Whilst designed to measure impact they are also metrics of performance. These, and other performance indicators, such as the time taken in the Digital Forensics Unit to examine the seized items in the average case, can be constructed from the Safenet dataset as it stands. There is scope to devise further bespoke Impact Points and associated Impact Metrics specifically for the purpose of the measurement of performance. For example, these could include those designed to allow the measurement of performance in areas such as victim experience and proportionality.

The repurposing and further development of Impact Metrics to produce performance indicators offers a number of useful prospects. These include, the ability to make interforce comparisons to identify best practice, the use of predictive modelling to aid operational planning and the development of a statistical process control (SPC) methodology.

At the heart of the last of these is the plotting of the variation in time of performance indicators to produce statistical process control (SPC) charts similar to the plot shown in Figure 4.1. These charts offer a number of benefits in operational management, which have been well explored by NHS Improvement. That organisation has published valuable guidance on the production and use of such charts (NHS Improvement 2021). Whilst their adoption is recommended, it must be noted that their success will be predicated on their being sufficient data. This might require data aggregation across time, geographical area and/or crime type, which might lead to a diminution of utility. A piece of work specifically aimed at the development, optimisation and rollout of these charts may well prove both necessary and beneficial.

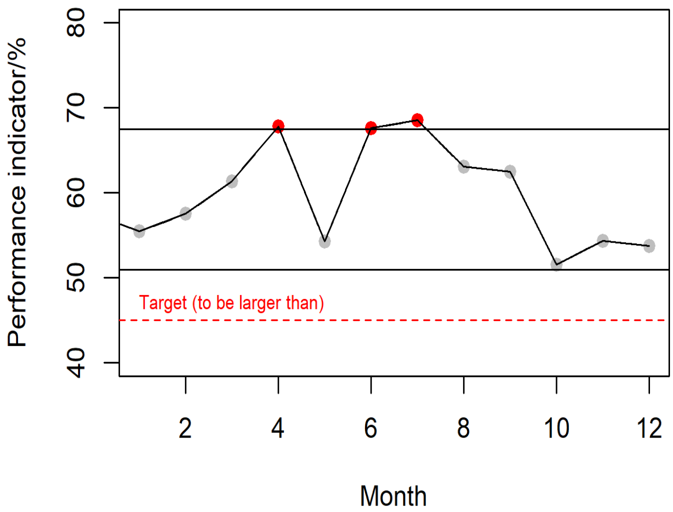


Figure 4.1 Example statistical process control chart

Based on the evaluation given above, the following recommendations are therefore made:

- The development and use of the following tools, all to be based on the IP approach:
    - regularly updated interactive dashboards;
    - inter-force comparisons to find best practice;
    - a statistical process control methodology;
    - the development of new performance indicators as needs be;
    - predictive modelling to aid operational planning.
- The provision of funding, as needs be, to allow the adoption of the above.

4.2.3 Further development facilitate forensic value optimisation and strategic decision making

It is clear from the Safenet study, reported here, that the Impact Point approach can be used to generate Impact Metrics that not only measure forensic value but which are also sensitive to changes in operational practice. This means that they offer the prospect of facilitating the maximisation of that value via studies that seek to understand the factors that cause improvements to the impact of forensic science.

It is, of course, entirely possible that factors that cause improvement in one area could also cause unwanted side effects elsewhere. It is important to, where possible, anticipate and measure these side effects before any strategic policy change is made. Where this cannot be done, the careful monitoring of performance after any such change, via tools such as statistical process control (see Section 4.2.2), can be used to identify and manage any problems that they cause.

The understanding created by these approaches can be used to inform strategic decision making to optimize forensic value.

Given the logic set out above, and the lessons learnt during the Safenet study, the following recommendations are made:

- All future work should include, by design, clear methods for distinguishing between causal associations and common cause effects, and the routine use of statistical tests to help detect effects that are due to chance alone. Where possible, these methods should seek to understand the mechanism(s) that drive any causal associations found.
- The widespread adoption of the Impact Point approach is recommended, as this will provide:
    - granular evidence of the impact of forensic science;
    - a tool for the optimisation of forensic value via the identification of patterns, trends and best practice, the design of pilot studies and effective policy implementation.
- Development of experimental and observational frameworks that include both statistical and qualitative methods.
- Further development and funding of the research commissioning process to stimulate innovation.

4.2.4 Transparency

Transparency in research brings many benefits. These include the promotion of further knowledge creation and innovation. The following recommendations are therefore made:

- Publish the findings of this and related studies via peer reviewed journals.
- As far as practicable, follow the tenants of reproducible research[50].
- Consider generation of datasets with greater levels of anonymisation or which are fully synthetic[51], allowing wider distribution. This will encourage and facilitate research and development, whilst enhancing transparency and promoting reproducible research.

4.3 Further recommendations concerning forensic methods, early triage use and data analysis

The study reported here has found strong evidence for the value of digital forensic methods in online child exploitation (CSE) cases. It has also found multiple associations between the use of early triage and a number of operational improvements. Although the cause of those improvements is not known, they are well explained by the ability of early triage to inform investigative decision making. Lessons have also been leant from the data analysis process that has been used in this study.

---

[50] For an introduction to reproducible research, see Alston, J M and Rick, J A (2021) A beginner's guide to conducting reproducible research, *Bulletin of the Ecological Society of America*, **102**(2) e01801 https://doi.org/10.1002/bes2.1801

[51] In this context, synthetic data are those which are fully computer generated. This generative process is informed by the properties of the observed data. Its purpose is to remove all records that could disclose identity whilst creating data that closely mimic the statistical behaviour of those observed data.

Based on the points made in the previous paragraph, the following recommendations are also made:

1. **Forensic methods**
    a. Continue the use of digital forensic methods in online CSE cases.
    b. In intelligence initiated online CSE cases, undue delay between the receipt of intelligence and scene attendance should be avoided. However, if such delay occurs, the resourcing of the examination of devices for digital evidence should not be reduced.
    c. Conduct research using the IP approach to establish:
        i. how the value of forensic science can be enhanced yet further. This could be within the field of online CSE cases or beyond. For example, a pilot study could be conducted to find the forensic value of trace evidence in the detection of cases of domestic abuse or of cases of burglary committed by prolific offenders;
        ii. the utility of early triage using digital forensic methods in priority crime categories beyond online CSE, such as Rape and Serious Sexual Offences (RASSO) and Serious and Organised Crime (SOC).
    d. Use the findings from 1.c to further boost the value of forensic science.
    e. Conduct research to better understand the impact of early triage (as defined in this study) on the workload of the Digital Forensics Unit.
2. **Early triage in online CSE cases**
    a. Continue the use of early triage where it is currently used. If the timeliness of the early interview is of importance, such triage should be conducted at the scene, not elsewhere, provided that this option is available.
    b. Conduct research to:
        i. establish the cause of the operational enhancements that are associated with the choice to use early triage;
        ii. look for any unwanted side effects of that choice.
    c. Use the findings from 2.b to maximise the benefits associated with the use of early triage and roll these benefits out to future cases that, under current practice, would not benefit from them.
    d. In forces where early triage is not currently used in online CSE cases, conduct pilot studies to develop evidence-led policy on its future adoption.
3. **Data analysis**
    a. Extend Impact Metrics to include *Cost Benefit*.
    b. Make routine use of statistical analysis.
    c. Make routine the benchmarking against external measures of impact, such as CJS outcomes.
    d. Analyse those aspects of the Operation Safenet dataset that have not yet been explored.

5. Conclusions

The Safenet study, reported here, is part of the Home Office's Impact of Forensic Science project. It is one of six proofs of principle designed to test the utility of the Impact Point (IP) approach that is at the centre of that project (Barrett 2021a and 2021b).

Safenet is an intelligence-led operation of Staffordshire Police. It protects children via the detection of crimes of online child sexual exploitation (CSE). It makes substantial use of forensic science in the form of digital forensic methods. Within that context, the study reported here has:

- demonstrated that data can be gathered that allows the measurement of the *Effectiveness* and the estimation of the *Timeliness* of digital forensic methods across a number of IPs;
- established that the IP approach can measure the value of forensic science to the criminal justice system (CJS) and do so in greater resolution than has been achieved before;
- made effective use of the IP approach to demonstrate the substantial contribution that forensic science makes to Operation Safenet;
- provided, for the first time, benchmark data on forensic performance that can be used for the purposes of policy development and operational management.
- identified a number of choices that are under professional control that are associated with the enhanced performance of digital forensic methods;
- found that the choice to use early triage is associated with very noticeable and statistically significant improvements to both the *Effectiveness* and *Timeliness* of digital forensic methods across a range of IPs;
- established that the associations referred to in the previous bullet point can be well explained by early triage's ability to provide evidence of criminality at the crucial early stages of an investigation. However, it is noteworthy that, for good reason, the Safenet study is observational and not experimental. This means that it is not possible to rule out the possibility that these associations have a common cause. In other words, there may be an unknown driver that affects the both the probability of the decision to use early triage and the enhancements in the performance metrics that are associated with that decision;
- found that positive early triage, as opposed to negative early triage, is associated with an increased likelihood of early interview occurrence. This is important because admission of guilt can only happen during an interview;
- found that when early triage was used, a change in its location was associated with a change in the typical (median) time between scene attendance and early interview. When early triage happened at the scene as opposed to elsewhere (i.e., a police station or the police county headquarters), this time was 66 minutes (22%) shorter. The cause of this effect is not known. However, statistical testing suggests that it might not have happened by chance alone but that further work is needed to verify this finding;
- found no evidence strong enough to show that delay between the date of referral[52] and the date of scene attendance leads to loss of digital evidence.
- learnt lessons that could be used to:
    - inform the automation of the data collection process;
    - better understand cause and effect;
    - reduce the opportunity for unconscious contextual bias;
    - create tools for effective identification of best practice and trends, performance-monitoring and management, and evidence-based policy development

Please see the executive summary at the start of this document for a list of recommendations that arise from this study's findings.

[52] In this context, referral means the police's receipt of the intelligence report that alerted them to the possibility of a case involving online child sexual exploitation (CSE).

6. References

Ahern, E.C., Sadler, L.H., Lamb, M, Gariglietti, G. M. (2016) Wellbeing of professionals working with Suspected Victims of child sexual exploitation, *Child Abuse Review,* 26, (2) pp 130-140

Alston, J. M., Rick, J.A (2021) A beginner's guide to conducting reproducible research, *Bulletin of the Ecological Society of America,* 101(2), Article e01801, https://doi.org/10.1002/bes2.1801

Association of Chief Police Officers (ACPO), Forensic Science Service (FSS) (1996) Using Forensic Science Effectively.

Barrett, M (2021a) *Impact of Forensic Science Project: Phase 1 Report*, Home Office

Barrett, M (2021b) *Personal communication*, multiple emails.

Barrett, M and Varraich, J (2021) *Personal communication*, multiple emails.

Bashir, M. S., Khan, M. N.A., (2015) A triage framework for digital forensics, *Computer fraud and security,* 2015, (3), pp 8-18

Blakey. D (2000) Her Majesty's Inspectorate of Constabulary — under the microscope Thematic Inspection Report on Scientific and Technical Support, HMIC.

Brookman, F., Jones, H. (2019) Research Insight: The Role of Forensic Sciences and Technologies (FSTs) in Homicide Investigation in Britain. University of South Wales, Leverhulme Trust, pp. 1-2

Brookman, F., Jones, H., Williams, R., Fraser, J. (2020) Research Insight: The Use of CCTV during Homicide Investigations: Contributions, Challenges and Risks. University of South Wales, Leverhulme Trust, pp. 1-8

Cohen, J (1988) Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ, USA: Lawrence Erlbaum.

Cooper, S (2021) *Personal communication,* email.

Field, A, Miles, J and Field Z (2012) Discovering statistics using R. London, UK: SAGE.

Galarnyk, M (2018) Understanding boxplots, https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51, (accessed 20 August 2021)

Gentry, E. Soltys, M. (2019) SEAKER: A Mobile digital forensics triage device, *Procedia Computer Science,* 159, pp 1652-1661

Home Office (2016a) Statutory Definition of child sexual exploitation. Government Consultation. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/500097/HO_DfE_consultation_on_definition_of_child_sexual_exploitation_-_final.pdf (accessed 31 August 2021)

Home Office (2016b) Forensic Science Strategy: A national approach to forensic science delivery in the criminal justice system, Her Majesty's Stationery Office, Print ISBN 9781474129336 Web ISBN 9781474129343, https://www.gov.uk/government/publications/forensic-science-strategy (accessed 23 Aug 2021)

Home Office, APCC and NPCC (2019) *Forensics Review: Review of the provision of forensic science to the criminal justice system in England and Wales*, Crown Copyright, https://www.gov.uk/government/publications/joint-review-of-forensics-2018-and-implementation-plan (accessed 23 Aug 2021)

Hood, C (2021) *Personal communication*, email.

Horsman, G. (2020) The collectors ranking scale for 'at-scene' digital device triage, *Journal of forensic Science,* 66, (1), pp 179-189

Horsman, G. Laing, C. Vickers, P. (2014) A case-based reasoning method for locating evidence during digital forensic device triage, *Decision Support Systems,* 61, pp 69-78

Horvath. F., Meesig. R., (1996) The criminal investigation process and the role of forensic evidence: a review of empirical findings, *Journal of Forensic Science*., 41 (6) pp. 963-970

Jackson, A R W and Jackson, J M (2017) *Forensic Science* (4th edn), Pearson, UK

Jafari, M and Ansari-Pour, N (2019) Why, when and how to adjust your p values? *Cell Journal* 20(4): 604-607. Available from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6099145/ (accessed 28 April 2021).

Jung, S., Ennis, L., Stein, S., Choy, A. Hook, T (2013) Child pornography possessors: Comparisons and contracts with contact-and non-contact sex offenders, *Journal of sexual aggression,* 19, (3), pp 295-310

Kloess, J., Woodhams, J. Whittle, H., Grant, T. Hamilton-Giachris, C. E (2019) The Challenges of identifying and classifying child sexual abuse materials, *Sexual Abuse,* 31, (2), pp 173-196

Kloess, J. Woodhams, J. Hamilton-Giachritsis, C. (2021) The challenges of identifying and classifying child sexual exploitation material: Moving towards a more ecologically valid pilot study with digital forensic analysts, *Child abuse and neglect,* 118.

Lalani, N; Jimenez, R B and Yeap, B (2020) Understanding Propensity Score Analyses, *International Journal of Radiation Oncology, Biology, Physics*, 107, (3) pp. 404-7 https://doi.org/10.1016/j.ijrobp.2020.02.638 (accessed 29 August 2021)

Long, M, Alison, L, Tejeiro, R, Hendricks, E and Giles, S (2016) KIRAT: Law Enforcement's Prioritization Tool for Investigating Indecent Image Offenders, *Psychology, Public Policy, and Law* 22, (1) pp 12-21 http://dx.doi.org/10.1037/law0000069

Ludwig, A., Fraser, J (2014) Effective use of forensic science in volume crime investigations: Identifying recurring themes in the literature, *Science & Justice*, 54, (1), pp 81-88

McLaughlin. L (2021) The Impact of Forensic Science on the Criminal Justice System: A Review of Literature, Home Office report entitled 'Impact of Forensic Science project: Phase 1 Report'.

Millar, C (2021) *Personal communication*, multiple emails.

Mislan, R., Casey, E. Kessler, G. (2010) The growing need for on-scene triage of mobile devices, *Digital Investigation,* 6, (3-4), pp 112-124

Moston, S., Engelberg, T. (2011) The effect of evidence on the outcome of interviews with criminal suspects, *Police Practice and Research,* 12, (6), pp 518-526

Ngejene, C., Eloff, J., Sefara, T., Marivate, V. N (2021) Digital forensics supported by machine learning for the detection of online sexual predatory chats, *Forensic Science International: Digital Investigation,* 36.

NHS Improvement (2021) *Making data count* https://www.england.nhs.uk/publication/making-data-count/ (accessed 29 August 2021)

Nickson, M. Venter, H. (2015) Taxonomy of challenges for Digital forensics, *Journal of Forensic Science,* 60, (4), pp 885-893

Owens, J., Eakin, J. Hoffer, T. Muirhead, Y. Shelton, J. L.E (2016) Investigative aspects of crossover offending from a sample of FBI online child sexual exploitation cases, *Aggression and Violent Behaviour,* 30, pp 3-14

Peterson, J.L., Hickman, M.J., Strom, K.J. and Johnson, D.J. (2013). Effect of Forensic Evidence on Criminal Justice Case Processing. *Journal of Forensic Sciences*, 58, pp. 79-90

Peterson. J.L., Mihajlovic. S., Gilliland. M., (1984) Forensic Evidence and the Police: The Effects of Scientific Evidence on Criminal Investigations, National Institute of Justice, Washington DC.

Rogers, M.K., Goldman, J. Mislan, R., Wedge. T., Debrota, S., (2006) Computer forensics field triage process model, *The Journal of Digital forensics, security and law,* 1, pp 19-37

Roman, J.K., Reid, S.E., Chalfin, A.J. and Knight, C.R. (2009). The DNA field experiment: a randomized trial of the cost-effectiveness of using DNA to solve property crimes. *Journal of Experimental Criminology*, 5, pp. 345-369.

Roux, C., Talbot-Wright, B., Robertson, J., Crispino, F., Ribaux, O. (2015). The end of the (forensic science) world as we know it? The example of trace evidence. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, vol. 370(1674).

Shaw, A, Browne, A., (2013) A practical and robust approach to coping with large volumes of data submitted for digital forensic analysis, *Digital Investigation,* 10, (2), pp 116-128

Shiaeles, S., Chryssanthou, A. Katos, V. (2013) On scene triage open source forensic tool chests: Are they effective?, *Digital Investigation,* 10, (2), pp 99-115

Spiegelhalter, D (2019) The art of statistics: Learning from data. London UK: Pelican.

Stover, C (no date) Contingency Table. From MathWorld–A Wolfram Web Resource, created by Eric W. Weisstein. https://mathworld.wolfram.com/ContingencyTable.html (accessed 28 April 2021).

Touche Ross (1987), Review of Scientific Support for the Police, Home Office, London.

Uberlaker, D.H., Shamlou, A. and Kunkle, A. (2018). Contributions of forensic anthropology to positive scientific identification: a critical review. *Forensic Sciences Research*, vol. 4(1), pp. 45-50

Wickham, H (2014) Tidy data, *Journal of Statistical Software*, 59(10). Available from https://www.jstatsoft.org/article/view/v059i10 (accessed 21 May 2021).

Woodman, P.A., Spiranovic, C., Julian, R., Ballantyne, K.N. and Kelty, S.F. (2019). The impact of chemical trace evidence on justice outcomes: Exploring the additive value of forensic science disciplines. *Forensic Science International*, 307, pp. 110-121

Wüllenweber. S., Giles. S., (2021) The Effectiveness of Forensic Evidence in the Investigation of Volume Crime Scenes, *Science & Justice*, 61, (5), pp.542-554.

5. List of Appendices

File name
Appendix 1 Impact Points & Definitions.pdf
Appendix 2 Variable definitions.xlsx
Appendix 3 Drop-down menu levels.xlsx
Appendix 4 Locations of case characteristics in forensic management systems.xlsx
Appendix 5 The Safenet variables and key to Appendix 6.pdf
Appendix 6 Safenet data visualisation.pdf
Appendix 7 Timeliness rules.xlsx
Appendix 8 Early_triage_and_early_admission_of_guilt_V10.pdf