SAGE EMG-DfT Public Transport Risk Model

Data Protection Impact Assessment (DPIA) for use of CCTV data collected by Transport for London (TfL) and London Underground Ltd (LUL)

1. Need for a DPIA

Explain broadly what the project aims to achieve and what type of processing it involves. You may find it helpful to refer or link to other documents, such as a project proposal. Summarise why you identified the need for a DPIA.

Risk of transmission for COVID-19 is related of proximity to infected persons and contact with the virus deposited on surfaces. Globally, there are no comprehensive datasets that quantify the frequency, duration, or nature of these within public transport environments. Public transport is a high-risk environment owing to the high turnover of passengers and limited capacity.

This project intends to quantify successive surface contacts (e.g. with handrails) and proximity between people within public transport environments. The scope of the project includes vehicles and carriages, as well as stations, interchanges, and platforms. This document refers only to data collected by TfL/LUL, and is therefore limited to carriages, stations, and platforms.

The statistics generated will be used alongside the results of sampling to build a risk model. This will be used to inform the work of the Scientific Advisory Group for Emergencies during COVID-19, and potential government action or advice that results.

The primary means of generating statistics will be analysis of recorded video (CCTV) through a semi-automated system. Passengers and staff will be observed in these recordings. CCTV recordings are considered personal data, and a DPIA is required because we are monitoring publicly accessible areas on a large scale (under GDPR Art 35(3)(c)).

2. Description of the processing

Describe the nature of the processing: how will you collect, use, store and delete data? What is the source of the data? Will you be sharing data with anyone? You might find it useful to refer to a flow diagram or another way of describing data flows. What types of processing identified as likely high risk are involved?

Collection

TfL/LUL will provide historic CCTV recordings from a sample of their services. The sample will include; on-train footage for three different lines on a selection of days; and two underground stations' CCTV footage for the same days. The sample will be determined to reflect the presumed risk profile (i.e. skewed towards busier services). These recordings will be provided on TfL approved encrypted USB drives following TfL cyber security policies.

TfL operate surveillance cameras for a number of purposes. One of the purposes is to protect the health and safety of employees, customers and members of the public. TfL is sharing CCTV images for research and analysis purposes in support of this.

Use

The intended workflow is described below. Minor deviations from this may be required. The later stages for in-vehicle data are illustrated in Figure 1. Processing of station footage will be a simplified version of this, considering only the distance between individuals and not their contact with surfaces.

1) **Receipt of data:** Initial checks to confirm the received data is as expected and of usable quality. A small sample will be watched.

2) **Obscuration:** An algorithm will be applied to entire recordings to obscure any timestamps and blur any faces that can be recognised by a pre-trained computer vision model.

3) **Slicing and randomisation:** Recordings will be sliced into shorter segments and the order randomised, with the original timestamp data retained in a separate system available only to the project leads. These slices will be used for any manual counting that is required, or correction of the automated algorithms.

4) **Semi-automated contact detection and distance calculations**: A computer vision algorithm will generate candidates for potential contacts with surfaces. Short clips around this candidate will be reviewed (by a counting operator) through a secure interface specific to this project, to confirm whether the candidate involves genuine contact with the surface. The distance between passengers moving through queues, carriages or vehicles will be determined from bounding boxes associated with each person detected by the computer vision algorithm. All contacts and distances will be associated with a randomly assigned identifier for a passenger, only used to group data about an individual passenger during the scope of a single journey.

5) **Generation of statistics:** The anonymised contact and proximity data will be used to produce probability-density functions and other statistics that are aggregated beyond the level of individual journeys (e.g. a distribution covering all peak time bus services). Non-CCTV data may also assist in these statistics; these are not expected to contain any personal data. A separate DPIA will be produced for any non-CCTV data that comprises personal data.

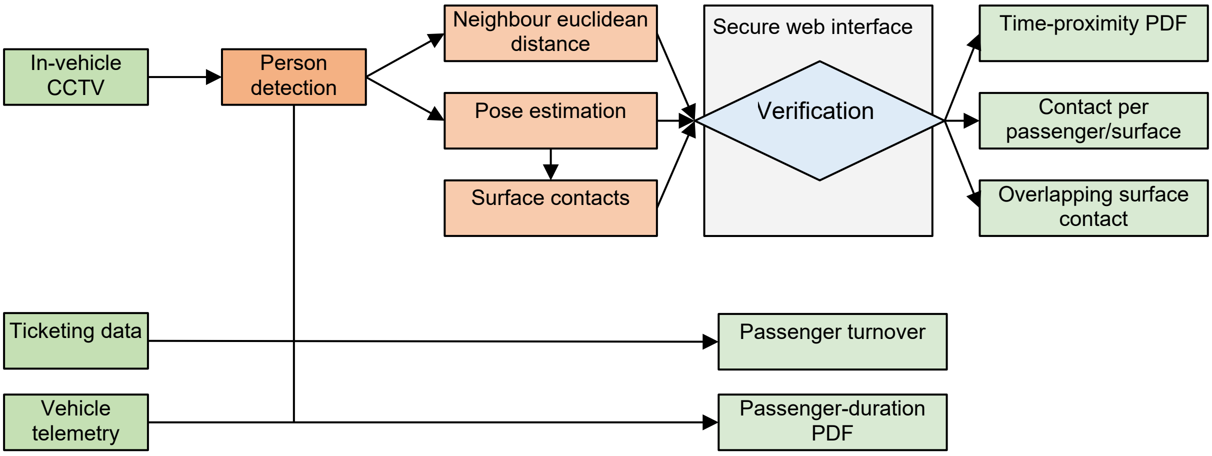

Figure 1 **Flow of data used to generate aggregate statistics for use in risk model.**

Storage

Raw recordings (as received)

Data will be transferred to off-network hardware encrypted (AES-XTS 256-bit) standalone drives with built-in keypads. TfL CCTV will be provided on approved encrypted USB drives following TfL cyber security policies sent by a courier service. The data will be transferred, and the USB drives will be wiped and returned to TfL.

To enable an automated data processing pipeline, to be run when changes are made to the computer vision model, raw data may also be transferred to a storage container in the cloud. This will be configured as described in Appendix 1 and be segregated from any other systems running in the cloud. The container itself will be encrypted and the files within it will be encrypted. Servers performing automated processing on this raw data will download individual videos from the cloud container, process it, and then immediately destroy the raw data.

Obscured and blurred recordings

Storage on a Newcastle University-administered web server backed by Isilon storage array.

Deletion

Raw recordings (as received)

Whichever is soonest: Not more than 12 months, two months following acceptance of the work for publication in a peer-reviewed academic journal.

Obscured and blurred recordings

Video data will be deleted once the obscured recordings have been processed by the algorithms and the counts verified by one or more counting operators (up to three counting operators may be required to independently count a section of video if there is a discrepancy between the first two).

Count data and aggregate statistics

These will be retained for research purposes for a minimum of 10 years (and published in aggregate) but will not contain personal data. Once the video data is destroyed the data will transition from pseudonymous to anonymous data, as we will have no means of identifying the individual. 10 years is the standard minimum retention period for research data, specified by UK Research and Innovation.

Describe the scope of the processing: what is the nature of the data, and does it include special category or criminal offence data? How much data will you be collecting and using? How often? How long will you keep it? How many individuals are affected? What geographical area does it cover?

Nature of data

The data is personal, as it captures the behaviours of individuals that could be identified. The obscured and blurred data is also personal data, as those with distinctive clothing could be readily identified or reidentified. We are not intending to capture any special category personal data or criminal offence data.

Quantity and frequency of data

Initially a one-off exercise, involving roughly ~500 hours across several cameras. The sample will include; on-train footage for three different lines on a selection of days; and two underground stations' CCTV footage for the same days. The exact quantity of data may vary but is limited by the manual effort required to obtain it. The exercise may be repeated (perhaps monthly) to provide longitudinal data until shortly after COVID-19 restrictions cease, as the gradual easing of restrictions is likely to change the transmission risk. TfL will keep a record of all footage provided to the research team.

Describe the context of the processing: what is the nature of your relationship with the individuals? How much control will they have? Would they expect you to use their data in this way? Do they include children or other vulnerable groups? Are there prior concerns over this type of processing or security flaws? Is it novel in any way? What is the current state of technology in this area? Are there any current issues of public concern that you should factor in? Are you signed up to any approved code of conduct or certification scheme (once any have been approved)?

Relationship with individuals, their control, and potential vulnerable groups

The research team have no direct connection to the individuals (i.e. passengers) within the video footage. Some of the individuals within the footage may belong to vulnerable groups. They are likely to have no control over whether CCTV is recorded in their presence, with the data being routinely collected for a range of lawful purposes.

Expectations of individuals

Passengers are likely to be aware of extensive CCTV in and around public transport networks. Signage is prominent as part of a crime prevention strategy and for data compliance. Most passengers are likely to be aware that CCTV may be used for purposes other than crime prevention, such as in addressing insurance claims and for compliance.

Transport for London's privacy page (https://tfl.gov.uk/corporate/privacy-and-cookies/cctv) informs passengers how TfL, including its operating subsidiaries, use personal data collected via CCTV across London's transport network. On a case by case basis we may use and share CCTV images for research and analysis purposes.

Technology concerns and maturity

CCTV is a well-established technology. We are applying computer vision and machine learning to the recorded footage to assist our work. We are not relying on computer-based decision-making processes with direct consequences for those appearing in the video footage (the risk model may result in guidance that changes practice).

Facial recognition is increasingly controversial. We will not use facial recognition algorithms. The cameras are generally of insufficient quality for these algorithms to work at distance, and our approach to blurring faces relies on an algorithm trained to detect the shape of a head rather than the features within it.

Codes of conduct, certifications, and other measures

This research is university-led. The university has strong policies in place concerning research ethics, and broad experience with handling sensitive data including classified government documents and sensitive category health data.

All research projects are required to conduct a risk assessment for ethics and data protection before commencing. This project was flagged as requiring additional review, because of the nature of the data required, and has subsequently been considered by Newcastle University's Faculty of Science, Agriculture and Engineering Ethics Committee. The committee approved the project based on the safeguards and measures described in this DPIA.

The Principal Investigator is a member of the Chartered Institute for IT and bound by its Code of Conduct.

Describe the purposes of the processing: what do you want to achieve? What is the intended effect on individuals? What are the benefits of the processing for you, and more broadly?

Through quantification of proximity and surface contacts on public transport, the subsequent risk modelling work will be used to make recommendations for measures that improve the safety of public transport. For example, this might include changes to the frequency and nature of cleaning, advice to passengers to remain seated until vehicles are stationary, or it may be possible to demonstrate that an increase in capacity would have minimal consequence for transmission risk.

There is no expectation that any of the data obtained or processed within this study will have direct consequences for individual passengers. It will not be used for enforcement of rules, shaming, or in unaggregated form by this research team.

3. Consultation process

Consider how to consult with relevant stakeholders: describe when and how you will seek individuals' views – or justify why it's not appropriate to do so. Who else do you need to involve within your organisation? Do you need to ask your processors to assist? Do you plan to consult information security experts, or any other experts?

Consultation directly with passengers that may appear in the recorded footage will not be undertaken and is considered disproportionate for this project. The quantity and geographic spread of footage to be analysed is low enough that the same individual would be highly unlikely to feature more than once within the analysis. The research team has no contact details or ready means of contacting those who will appear within the video footage, and the urgency of the work would make a more general consultation with passenger groups (for example) difficult to achieve. A widespread consultation at the locations involved in the study would also risk invoking a change in behaviour.

All transport operators involved in the study will be invited to comment on the approach described in this DPIA, and any measures they recommend to further mitigate risks without compromising the quality of data obtained will be seriously considered.

Newcastle University's ethical approval process considers the moral dilemmas associated with research, as part of its terms of reference. A senior lecturer in security within the School of Computing was consulted on ethics and technical measures.

4. Necessity and proportionality

Describe compliance and proportionality measures, in particular: what is your lawful basis for processing? Does the processing actually achieve your purpose? Is there another way to achieve the same outcome? How will you prevent function creep? How will you ensure data quality and data minimisation? What information will you give individuals? How will you help to support their rights? What measures do you take to ensure processors comply? How do you safeguard any international transfers?

Lawful basis for processing

Public task is our lawful basis for processing. Newcastle University is a public authority established by an act of parliament. Research is part of the university's purpose as described in its statutes. This research is being conducted in the public interest at the request of the UK government, in response to a significant health concern.

With regard to public task, the ICO specifically states that further "processing for certain purposes should be considered to be compatible with your original purpose," listing scientific and statistical purposes as examples of this.

Likelihood of achieving outcome, and alternative approaches

Alternative methods have been ruled out. Preliminary examination of camera images makes it highly likely that we will be able to obtain the required statistics using this approach. It is not feasible to observe contacts in person during the current environment, without endangering research staff. If that were not the case, the presence of a person counting surface contacts might affect behaviour and they would likely have a worse vantage point than elevated CCTV cameras.

Data quality, minimisation and preventing function creep

The hybrid approach of part-computer assisted counting and measurement is intended to ensure accuracy. Disparities between counts by two people will result in a third count by a different person. The intention is to guarantee a level of accuracy that prevents any need to repeat the work and observe additional recordings.

The total quantity of footage to be analysed has been capped based on capacity and current assumptions about obtaining a representative sample. This quantity may need to be amended based on preliminary results.

Data is minimised through the removal of features that might allow correlation against third-party data (e.g. timestamps over the video) and faces are blurred because they are not required for the analysis. Clothing and other features cannot be blurred without degrading the study.

Data will only be used for the purposes of this study and no additional use will be made outside of the scope of the purposes stated for this study.

Supporting the rights of individuals

We will respond to subject access requests as required, and if objections are raised, we will consider whether our methodology can be amended to address their concerns without a detrimental effect on the data we produce.

Operators and data providers will be asked if they are content with their involvement in this work being publicised before, during, after or never. As a university, we may nonetheless have to disclose details of this work under the Freedom of Information Act or Environmental Information Regulations.

A description of the work will be placed on a public website, including contact details for additional information or to exercise data protection rights.

5. Risk assessment and reduction

Potential source of risk and nature of potential impact on individuals. Include associated compliance and corporate risks as necessary.	**Likelihood of harm** (remote, possible, or probable)	**Severity of harm** (minimal, significant, or severe)	**Overall risk** (low, medium, or high)	**Measures to reduce or eliminate risk**	**Effect on risk** (eliminated, reduced, or accepted)	**Residual risk** (low, medium, or high)
Public objections to use of data for this purpose Including those reported directly and those expressed on social media or elsewhere. False information could precipitate mistrust in public transport operators, fear of being watched by passengers, staff believing the work may be used for disciplinary or enforcement activity, etc.	Possible	Minimal	Low	Total volume of recorded footage involved is small, highly unlikely that the same person would appear more than once in footage obtained. Sampling strategy to avoid repeat counting on the same journeys over different days. Website describing the work and safeguards in place.	Reduced	Low
Theft or loss of physical media Including during transit, from university premises, or from the research team while working from home. Risk of data becoming available to potentially hostile parties, and reputational damage for organisations involved. Unauthorised persons could view and share footage.	Possible	Significant	Medium	All physical media to be encrypted, using hardware-level encryption on specialised drives that are certified for use with classified data. TfL will be providing the raw CCTV footage on encrypted USB drive sent via courier service.	Reduced	Low
Theft, hack, or compromise of server with data Including external actors and students and staff at the university, or unintentional misconfiguration. Risk of access to results and blurred footage, and potential further dissemination.	Possible	Significant	Medium	Server to be used is part of a university centrally administered cluster that includes regular scanning for malware and vulnerabilities and sits behind the university firewall that can detect common attack patterns. Only blurred and obscured footage will be stored on these servers. Data that would allow the blurred recordings to be associated back to a specific timeframe will be stored on a different system. Named-only access to limited members of the project team for development and deployment access on the server. Two-factor authentication for counting operators and software engineers, on both SSH connections and HTTPS connections.	Reduced	Low

Potential source of risk and nature of potential impact on individuals. Include associated compliance and corporate risks as necessary.	**Likelihood of harm** (remote, possible, or probable)	**Severity of harm** (minimal, significant, or severe)	**Overall risk** (low, medium, or high)	**Measures to reduce or eliminate risk**	**Effect on risk** (eliminated, reduced, or accepted)	**Residual risk** (low, medium, or high)
				Encrypted connections (HTTPS) and a restricted number of ports accessible from a restricted range of IP addresses.		
Intentional or unintentional leak of personal data by counting operator Including viewing of the screen during work by those unauthorised, posting of images on social media, or access to an operator's workstation by those unauthorised.	Remote	Significant	Low	Counting operators and research staff asked to sign an agreement that sets out expectations and restrictions on how they perform their work, including confidentiality within their home-working environment. All counting operators will fall under the university's requirements on research ethics and could be subject to disciplinary action if they disclose personal information. Timeout within the counting interface that locks the system if the user has not moved the mouse or used the keyboard recently.	Reduced	Low
Unintentionally obtaining special category data Behaviours observed may be an indicator for special category data such as health, sexual orientation, race, etc. Risk of an operator being able to attribute a special category data characteristic to an individual.	Remote	Minimal	Low	Counting operators will be able to flag footage that is inappropriate for use in this research, allowing it to be excluded from the study altogether. Special category data is not actively being collected within the data generated by counting operators or the automated algorithms, and it is improbable that these behaviours would be observed with the sample skewed towards peak time journeys. Blurring in the footage makes it highly unlikely that the behaviour could be attributed to an individual, though a remote possibility remains.	Reduced	Low
Counting operator witnesses crime or injury Footage may capture events that could cause alarm or distress to the research team.	Possible	Significant	Medium	TfL LU CCTV manager will review if any incidents were reported during the time period if so this footage will not be provided.	Reduced	Low

Potential source of risk and nature of potential impact on individuals. Include associated compliance and corporate risks as necessary.	**Likelihood of harm** (remote, possible, or probable)	**Severity of harm** (minimal, significant, or severe)	**Overall risk** (low, medium, or high)	**Measures to reduce or eliminate risk**	**Effect on risk** (eliminated, reduced, or accepted)	**Residual risk** (low, medium, or high)
Risk to the mental health of the research team.				Data providers and operators will be asked to exclude any footage they know relates to an incident. Any remaining incidents are likely to be minor. Counting operators will be instructed to flag footage that they are uncomfortable watching, such as involving drunken behaviour or harassment, and the footage will be excluded from the study. Any incidents should also be flagged to TfL to ensure appropriate action is taken.		
Identification or reidentification of persons within footage Precipitated through deficiencies in the blurring algorithm that may not identify all faces, or through the combined effects of other data or personal knowledge (e.g. friends' clothing). Risk that when combined with other knowledge, someone may be able to infer additional information about a journey such as its purpose (e.g. a person recognising their own partner may suspect an affair).	Possible	Minimal	Low	Counting operators will be able to flag footage where the blurring algorithm has failed significantly, such as not blurring any of a specific person's face for a prolonged period. The algorithm parameters would then be tweaked, and the blurring algorithm reapplied. Confidentiality and obscuration associated with the research project makes it unlikely that harm would result even if a person were recognised within the footage. As an example, if someone recognised their own partner, not having access to when the footage was recorded provides an additional safeguard against being able to infer the purpose of the journey, though does not fully eliminate the risk. Counting operators and the research team will be instructed to stop watching any footage if they recognise any of the persons within it.	Reduced	Low
Project scope or function creep The algorithms developed to semi-automate this approach could potentially be repurposed, and ultimately used in enforcement action or for automated decision-making. Other research projects may want to use the data. Minor amendments may be required to the project methodology to achieve accurate results.	Remote	Significant	Low	Tight constraints on who can access the data, retention of the raw data away from places where it can be more generally accessed, and clear agreements with the data providers on the purpose of the data provision. Ethical approval required for all new research projects within the university, which would require consent	Reduced	Low

Potential source of risk and nature of potential impact on individuals. Include associated compliance and corporate risks as necessary.	**Likelihood of harm** (remote, possible, or probable)	**Severity of harm** (minimal, significant, or severe)	**Overall risk** (low, medium, or high)	**Measures to reduce or eliminate risk**	**Effect on risk** (eliminated, reduced, or accepted)	**Residual risk** (low, medium, or high)
Risk that the project encroaches into collecting special category data or falls outside of the lawful purposes for which the data was originally collected (i.e. public health/safety).				from the data providers for an additional study. Data will not be transferred outside of the university unless compelled by law.		

6. Sign off and outcomes

Item	Name/date	Notes
Measures approved by:	Maureen Wilkinson (NU DPO) Dr Luke Smith (NU PI)	Integrate actions back into project plan, with date and responsibility for completion.
Residual risks approved by:	Maureen Wilkinson (NU DPO) Dr Luke Smith (NU PI)	If accepting any residual high risk, consult the ICO before going ahead.
Ethical approval date and ID:	20-SMI-017 Granted 03 June 2020	Ethics committee will consider GDPR compliance, data protection and broader matters relating to internal standards, funder requirements, university ethos, and reputation.
Summary of ethics panel advice: Project approved with condition that identifying information is protected using every measure possible. The panel are aware that no algorithm is perfect, and that there will be instances where faces are not blurred by the algorithms.		
Consultees at data provider:	Rita Scollan, Privacy Adviser Review by: Simon Guild, Head of Privacy and Data Protection Richard Bevins, Head of Information Governance & Data Protection Officer	
Summary of consultees' comments: Amendments to reflect TfL's internal policies and the mechanisms through which data will be shared. Discussion of whether raw data may be stored in the cloud to allow a fully automated data pipeline for processing video. Consideration needed for whether individuals' expectations have changed since COVID-19. Inappropriate or concerning behaviour witnessed should also be reported back to TfL. Further documentation, completion of data protection training and an Academic Confidentiality Agreement have been provided and approved to ensure appropriate processing and security is applied to the CCTV data.		
DPIA will be kept under review by:	Dr Luke Smith Newcastle University	Next review: December 2020

Appendix 1 Details for use of Azure storage

Where storage using Microsoft Azure is involved for personal data, it must be configured as follows. Any updates to their best practice guidance should also be considered.

Consideration	Requirements
Location	Within the UK only. No transfers to regions outside of the UK.
Storage account	Role-based access control (RBAC). Auditing enabled. Logging enabled for how requests are authorised. Resource manager-based deployment and governance, access to managed identities, access to Azure Key Vault for secrets, and Azure AD-based authentication and authorisation for access to all Azure storage data and resources.
Storage configuration	Enable the 'secure transfer required' option (HTTPS transfers only, HTTP will fail). AES-256 encryption required for files in storage. Soft delete disabled. Files removed in line with DPIA and agreed retention policies.
Threat protection	Enable the 'advanced threat protection for Azure Storage' option.
Access to blob data	Require clients using shared access signatures (SAS tokens) to use HTTPS when accessing blob data.
Network configuration	Access to the storage account should be restricted to individual IP addresses or small ranges.