13th March 2014

Christopher Graham,
Information Commissioner,
Wycliffe House, Water Lane,
WILMSLOW, Cheshire SK9 5AF

Dear Chris,

Complaint: NHS Data Storage in the Google Cloud

We are writing about recent disclosures of the use of NHS data by PA consulting and we request that your office investigate apparently serious breaches of the Data Protection Act 1998.

Background

As part of a data analytics project, the NHS Information Centre (NHS IC) – a predecessor of the Health & Social Care Information Centre (HSCIC) – entered into an agreement to share Hospital Episode Statistics (HES) data with PA Consulting Group (PA) in November 2011. The data sharing agreement allegedly imposes a number of restrictions on PA's use of the HES data, including a limitation on the number of people that can access the data, a restriction on sharing the data with third parties, and an obligation to erase the data following the termination of the agreement.

According to an HSCIC press statement, the shared datasets include "pseudonymised" HES on all NHS inpatient treatments, outpatient appointments and A&E attendances in England.[1] Each HES record generally contains a broad range of information about individual NHS patients, such as age group, gender and ethnicity, diagnostic and treatment codes, and information about the

[1] *HSCIC Statement: Use of data by PA consulting*, 3 March 2014, available at: http://www.hscic.gov.uk/article/3948/Statement-Use-of-data-by-PA-consulting.

[2] See, HSCIC, *What HES data are available?*, available at: http://www.hscic.gov.uk/hesdata.

location where the patient was treated and where he/she lives.[2] By default HES data contain the patient's postcode and date of birth, which in combination are enough to re-identify about 98% of patients; it is unclear whether these data were redacted in this case. Even without these data, longitudinal medical records are often very easy to re-identify.

In order to analyse and manipulate the HES data, PA decided to use third-party tools supplied by Google. Specifically, PA uploaded the HES data to Google Storage, and processed it via a Google analytics service, Google BigQuery. (Google BigQuery is a cloud service that allows interactive analysis of large data sets.) While little is known about the agreement between PA and Google, PA did provide NHS IC with a written confirmation that no Google staff would gain access to the HES data and that "access continued to be restricted to the individuals named in the data sharing agreement."[3] Neither PA nor HSCIC have provided any information about the assurances, if any, they received from Google. It is difficult to see how PA could exclude the possibility that Google engineers might access the data, whether of their own volition or pursuant to a lawful access request from a US government agency, and this raises the question of whether PA's confirmation was anything more than just wishful thinking or a desperate attempt at blame avoidance.

When the details of this data-sharing arrangement became public, stakeholders were highly concerned. MP Sarah Wollaston, who sits on the Health Select Committee, tweeted: "So HES data uploaded to 'Google's immense army of servers', who consented to that @hscic?"[4]. This concern is unsurprising given Google's record on privacy; in recent years, Google was found to have breached EU data protection law by the EU's Article 29 Working Party, as well as by regulators in a number of Member States.

Issues

In respect of those HES records that qualify as personal health information, a range of complex legal and professional obligations restrict or prohibit the use and disclosure of such data, including the UK Data Protection Act 1998, the common-law duty of confidence, the Human

[2] See, HSCIC, *What HES data are available?*, available at: http://www.hscic.gov.uk/hesdata.

[3] *HSCIC Statement, supra* n. 1.

[4] https://twitter.com/drwollastonmp/status/440275592655949824.

Rights Act 1998, the NHS Confidentiality Code of Practice, and the Information Security NHS Code of Practice.[5]

Although PA's press statement claims that the shared dataset does not contain any information that could be linked a specific individual,[6] it is quite unclear how that statement could be correct. Even if the HES dataset stored in Google's cloud services does not contain a patient's name or NHS number, the data there may be easy to link to a specific individual and hence will often constitute sensitive personal data. A record of a catheter ablation procedure at Hammersmith Hospital on October 19th 2003 can be linked with high probability to Tony Blair on the basis of press reports of his treatment for atrial fibrillation, and if the dataset permits episodes relating to him to be linked, then sensitive personal information relating to his other treatment episodes may be very easy to find. A large research literature going back to the late 1970s explores the substantial risk that individuals may be re-identified from pseudonymised datasets. [7] The data sent to the Google Cloud must therefore be treated as personal data, and indeed as sensitive personal data, for the purposes of European and UK data protection law – even if postcodes and dates of birth were in fact removed. We note that neither HSCIC nor PA has so far claimed that postcodes were removed.

We request that you conduct an investigation to determine whether the personal health information of NHS patients, including the signatories to this letter, was uploaded to Google systems.

If so, storing and processing such data would probably breach numerous rules and regulations. In particular:

- Personal health information should not be disclosed to third parties except in very limited circumstances. The data-sharing agreement between NHS IC and PA restricts the number of individuals who can have access to the HES data; PA has made a specific commitment to NHS IC not to allow Google staff to access the data. Yet it is unclear that they got adequate assurances from Google.

[5] The UK Department of Health has developed an online Information Governance Toolkit (IGT) that consolidates all applicable legal rules and central DoH guidance as a set of information governance (IG) requirements. The IGT enables NHS organisations and third parties providing services to NHS organizations to assess their compliance with current legislation, government policy and national guidance.

[6] *PA Consulting Group statement: use of HSCIC data,* 3 March 2014, available at: http://www.paconsulting.com/introducing-pas-media-site/releases/pa-consulting-group-statement-3-march-2014/.

[7] It has been clearly established (and has long since been known amongst academics, researchers and practitioners) that such minimal§ "de-identification" does not prevent data from large databases from being re-identifiable.

- The purposes for which personal information of NHS patients can be used are restricted. As a general rule, unless there is a legal basis for the use of data for other purposes (e.g., patient's express consent), personal information of patients may only be used to provide care services and for related purposes (e.g., to improve the quality of healthcare management or service delivery). In particular, the use of patient health information for commercial purposes, including the provision of advertising, is prohibited. But Google's cloud-service agreements allow Google to process customers' data for open-ended and vague purposes, which leaves open the possibility that Google may be processing personal health information for its commercial benefit and in particular to optimise the provision of advertising.

- Detailed security standards apply to the processing and storage of health information. Among other obligations, the UK Department of Health (DoH) has published detailed guidance on suitable encryption algorithms for NHS patient data.[8] It is unclear that the security measures Google applies to its cloud services are compliant. We refer you in particular to recent disclosures by Edward Snowden to the effect that foreign intelligence agencies were routinely harvesting personal information of Google customers on the unencrypted backbone links between its data centres, and that GCHQ did not insist on minimisation of personal information of UK citizens within 5 eyes (unlike the CSE which insisted on such minimisation for Canadian citizens).

- The transfer of NHS patients' personal information outside the UK is heavily restricted. In particular, the DoH guidance makes clear that such information must not be transferred outside the UK unless an appropriate assessment of risk has been undertaken and appropriate controls implemented; the transfer is notified to your office; the decision to transfer the data has been taken by a senior manager with the required authority; an assurance statement is obtained from third parties that process the data overseas; and – in most cases – the patients to whom the data relates have been notified about the transfer. As Google has no data centres in the UK, and takes the position that its customers' data may be stored in any of its datacentres[9], managers contemplating the use of Google services for personal health information should have properly followed the procedure for sending such information overseas.

---

[8] See, NHS Information Governance, *Guidelines on Use of Encryption to Protect Person Identifiable and Sensitive Information*, 2008, available at: http://systems.hscic.gov.uk/infogov/security/encryptionguide.pdf

[9] See, IT News, *Google: Who cares where your data is?*, 9 June 2011, quoting Chief security officer for Google Apps, Eran Feigenbaum, available at: http://www.itnews.com.au/News/260041,google-who-cares-where-your-data-is.aspx.

- Personal health information must be deleted when it is no longer required for a specific purpose. This commitment has apparently been repeated in the data sharing agreement between NHS IC and PA, so that PA is supposed to delete the HES data once the agreement terminates. But it is unclear that Google is subject to similar restrictions. Indeed, in the past Google has failed to provide strong commitments to its cloud customers to delete data during provision and after termination of the service.

The storage of large amounts of sensitive personal health information in a US cloud service is particularly concerning because of the precedent it may set. Google may advertise a motto of 'don't be evil' and some of us individually may be prepared to accept assurances from them (one of us – Anderson – is a former Google employee). However not all UK data subjects will be prepared to accept such assurances – not everyone uses gmail. Furthermore, there are many other service providers with a range of corporate cultures. Some overseas service providers are very much less trustworthy, and fall completely outside your regulatory scope as they have no UK presence; we are concerned that our personal health information will end up there next. Yet this need not happen; there are many UK and EU service providers who fall completely within the scope of the Data Protection Directive, and we note that even Microsoft will now store personal data in the EU if customers demand it.

Questions

We request that you investigate the potential breaches of UK laws and regulations resulting from the uploading of patient data to Google's cloud services. This relates not just to the Data Protection Act 1998 directly, but to the relevant NHS regulations and the relevant human-rights law (including I v Finland) as these all set the reasonable expectations that patients had when they supplied their information to the NHS, and thus are fundamental for fair processing.

Among the questions that must be asked:

- Precisely which patient data were stored outside the UK? Did they relate to single episodes or linked records? Did they contain postcode, date of birth, NHS number, or a pseudonym such an encrypted NHS number? The statements from PA and HSCIC deny that a name or full address was included, and PA denied there was a full date of birth. Neither has denied postcode, or year of birth, or the use of a pseudonym that would enable episode records to be linked. HSCIC mentions 'pseudonymised' data, which suggests a pseudonym. We as patients and data subjects (as well as advocates) would like to know the details.

- What kind of privacy risk assessment was carried out by PA and NHS IC prior to deciding to store, or to consent to the storage of, the data in Google's cloud services?

- If data were transferred under Safe Harbor (as one might expect), the Controller still needs an Art.17 contract governing security of processing. Does this contract exist, and if so, have its adequacy and lawfulness been verified? Can we see it?

- How are HES data protected against access by unauthorised parties, including Google engineers? Were any encryption methods used to protect the data (other than the TLS encryption used to protect the link from the client to the Google Front End) and who has access to the encryption keys?

- What assurances were obtained that the HES data could only be used for healthcare purposes? In particular, has Google made any commitments not to use the data for its own commercial purposes, such as targeting adverts or analytics?

- As the data were transferred to servers outside the UK, have the requirements under the Data Protection Act 1998 and the DoH guidance been complied with?

- What measures have the parties taken to ensure that the HES data cannot be accessed by foreign government agencies using their local powers, rather than having to go through UK lawful-access procedures?

- Were adequate arrangements made to ensure that Google's data processing activities can be audited?

- Has the specific commitment to erase the HES data once the data sharing agreement terminates been extended to Google?

We ask you to investigate these issues as a matter of urgency.

Yours sincerely,

Ross Anderson
Chair,
Foundation for Information Policy Research
Ross.Anderson@cl.cam.ac.uk

Phil Booth
Coordinator,
medConfidential
phil@medconfidential.org

Nick Pickles
Director,
Big Brother Watch
Nick.Pickles@bigbrotherwatch.org.uk