The NHS as a proving ground for cryptosystems

Chris R. Dalton
Trusted Systems Laboratory
HP Laboratories Bristol

HPL-2003-203
October 3rd , 2003*

healthcare, identity-based encryption, PKI

This paper examines the challenges the National Health Service poses as an environment for public-key cryptography systems.

The NHS is Europe's largest single employer with over 1.2 million staff. It provides lifetime healthcare for most of its population, and has done so for fifty-five years. In the last decade, it has launched several major programmes to develop NHS-wide information systems.

Just the scale of the NHS is daunting. But systems handling patients' medical records are subject to a plethora of laws, policies, guidelines, and practices for controlling the access, use, and storage of the information. Taken together, the rules are complex and sometimes contradictory, and in addition, have to be balanced against both patients' express wishes and clinical needs.

Cryptography is an obvious means to secure and protect confidential information. Recently, identity-based public key cryptography schemes not only seem easier to deploy than previous schemes, but also seem equal to the challenges. In this paper, we give a detailed overview of the features and challenges the NHS environment presents to uses of cryptography, to qualify our impressions of our cryptosystem and to guide our future efforts to develop it.

The NHS as a proving ground for cryptosystems

Chris R Dalton
Trusted Systems Laboratory
Hewlett-Packard Laboratories, Bristol

This paper examines the challenges the National Health Service poses as an environment for public-key cryptography systems.

The NHS is Europe's largest single employer with over 1.2 million staff. It provides lifetime healthcare for most of its population, and has done so for fifty-five years. In the last decade, it has launched several major programmes to develop NHS-wide information systems.

Just the scale of the NHS is daunting. But systems handling patients' medical records are subject to a plethora of laws, policies, guidelines, and practices for controlling the access, use, and storage of the information. Taken together, the rules are complex and sometimes contradictory, and in addition, have to be balanced against both patients' express wishes and clinical needs.

Cryptography is an obvious means to secure and protect confidential information. Recently, identity-based public key cryptography schemes not only seem easier to deploy than previous schemes, but also seem equal to the challenges. In this paper, we give a detailed overview of the features and challenges the NHS environment presents to uses of cryptography, to qualify our impressions of our cryptosystem and to guide our future efforts to develop it

Keywords: healthcare, identity based encryption, PKI

1 Introduction

1.1 The National Health Service

In 1948, the UK government launched a programme of unprecedented ambition and scale: to provide medical services to everyone, free of charge, from conception to death. Today, the National Health Service has an annual budget of more than GBP 61 billion (100 billion dollars) and, with over 1.2 million employees, is the largest single employer in Europe. Last year, its hospitals treated over 30 million patients, one-third of them as in-patients; general practitioners across the country saw an average of 1 million patients each day[33].

Throughout its life, the only constant has been change. The organization has had several major reorganizations of its structure and business model, brought on both by growth and by changes in the economic and political climate. The population it serves has grown by over a fifth, and – partly because of improved healthcare – the average age of patients has increased out of proportion. On top of that, the past five decades have seen phenomenal changes in medicine and technology: what was unimaginable in 1948 has become not only possible, but in many cases is now expected treatment.

In the mid-1990s, the UK Government announced plans to create a national healthcare network, with the aim of establishing NHS-wide medical records system and improving the efficiency of the healthcare services generally. In 1999, the NHS executive established the NHS Information Authority to oversee and coordinate the creation of "joined-up" NHS services, including a national electronic patient records system, and systems to handle booking, prescriptions, and pathology reports, among others.

1.2 Motivation

In 2001, a new public key cryptography scheme was developed based on the mathematics of pairings[5][10]. In HP Laboratories' Trusted Systems Lab, we, like others, were intrigued by possibilities the new scheme promised[32], and began to consider how we might apply some of them[9]. Secure email was an obvious candidate, we began work on a prototype for a particular messaging application in the NHS[7].

When we started to build the messaging prototype, it was largely instinct telling us the new cryptosystem was a good fit. One year on, it seems a good time to take a step back, look at the bigger picture, and try to find a way to qualify our instincts. What are the true problems that need to be solved? What problems does the NHS pose that we hadn't considered?

The purpose of this paper is to summarise the practical issues that someone designing or applying an cryptographic system for use in the NHS is likely to encounter. We begin with an overview of the NHS as an environment for information systems. We consider the sensitive information the NHS keeps about patients and the ways in which it processes it, and we look at the complex fabrics of rules and policies that govern access and processing of that information. With that in mind, we examine the implications for using public key cryptography as a component in NHS systems. Finally, we re-examine conventional and identifier-based PKC schemes in this light.

2 A Challenging Environment

In discussing information and cryptographic systems, we sometimes talk somewhat abstractly about scalability, of managing credentials and certificates, and so on. Designing information systems for the NHS, one faces some formidable challenges: the size of the organisation; the number of staff; the kinds of jobs they do; the sensitive nature of the information to process; and the range of time scales involved, from heartbeats to lifetimes.

We begin this section by considering the data in the system that is of most interest to us – information about patients – and how it is used. Then, in the largest part of the section, we look at the tangle of laws, rules, and policies that govern access to this information. Finally, we look at the NHS itself, its size, structure, and the implications they have for system design.

2.1 Patient Medical Records

Knowing the medical history of a person – the illnesses, symptoms and conditions they have had and the treatments they have received – is crucial to diagnosis and to safe, effective treatment. These pieces of information are among most sensitive and private that one person entrusts to another. No surprise, then, that confidentiality has been at the core of physicians' ethics since Hippocrates:

> "And whatsoever I shall see or hear in the course of my profession, as well as outside my profession in my intercourse with men, if it be what should not be published abroad, I will never divulge, holding such things to be holy secrets".

Yet the value of records goes far beyond the treatment of individuals: medical research, public health, and the planning and provisioning social and healthcare services are just a few of the important uses. It's also important to note the possible value to others, for example, to potential employers or insurers wishing to screen “high risk” applicants[1][2].

2.1.1 Patient Identifiable Information

Most discussions and documents about medical records distinguish between the clinical information in the notes, and the information that could potentially allow someone to identify an individual, either directly or indirectly.

Patient Identifiable Information (PII) includes names, addresses, identification numbers, pictures, sound recordings, and medical images. It also includes other less obvious information, such as relatively rare symptoms, diseases, or treatments, or small populations in statistical samples. Thus, identifying PII is not always a simple question of isolating certain fields in a record.

2.1.2 How Records are Used

The ways the people in the NHS handle patient information falls into two main categories. First, they store the information, occasionally reading it and less frequently, updating it, usually by appending. Second, they exchange it with others in the organisation. In either case, the confidentiality of the information is a prime concern.

A third category, bulk analysis of the data, is not yet in common use, but this is expected to change with the introduction of wide-scale digital systems.

2.1.2.1 Storage

Until recently, there was no practical option for storing medical records other than on paper. Apart from its longevity, paper also imposes a practical limit on how many records one person can process in a day. While this can be a drudge for clinical and administrative staff, at the same time it limits the possibilities for misuse.

Usually, records are kept at the place where the corresponding treatment took place. Thus, one patient's medical “record” might actually comprise the notes the GP keeps, a set in each of several hospitals that treated the patient for a recurring condition, and a set in a hospital at the other end of the country that treated an injury from a holiday accident. Most records pertain to one patient, although there are a few exceptions, notably maternity and neonatal records.

Retention periods

Information in a patient's record needs to be retained for some time for clinical and legal reasons. However, particularly with paper records, the NHS has had to balance this against the costs and practicalities of storage.

The guidelines for retention[14][23] are complicated. They depend on the patient, the treatment, any special circumstances, and ultimately on clinical advice and judgement.

For example, most records for adults are kept for at least eight years from the completion of treatment. They're kept longer if the treatment was part of a clinical trial (15 years), or for some conditions including transplants (11 years) and mental illness (20). Children's records are kept at least until their subjects reach the age of 25; similarly, maternity records are kept for 25 years. For some patients, such as members of the armed forces, records are held indefinitely.

Under these rules, records can outlive not only the people who created them, but also the institutions where they were first collected; they can also be older than the people who need to read them. Medical information systems need to be able to handle this.

2.1.2.2 Messaging and transport

Coordinating the various parts of the patient's history has so far been handled by simple protocols – a set of standard letters that carry pertinent information from one place to another. One form is used to tell a patient's normal doctor about emergency treatment their patient received at hospitals, clinics, or another practice. A “referral” letter is a request from a GP asking a specialist, often at a hospital, to examine and perhaps treat a patient; it contains the relevant symptoms and facts from the patient's history. The converse is the “discharge letter” from a consultant, which tells the GP what treatment the patient received, the prognosis, and recommendations for future care.

These letters are usually addressed to roles or to groups (departments), not to individuals. In many units, the so-called “Dear Doctor” letter is an explicit part of the referral protocol:

> Unless you need the patient to be seen by a particular consultant please address your Dear Doctor letter to the relevant department, rather than to a named consultant. This will help to expedite your referral.[29]

In any event, the people sending the messages often don't know the names of the people who will be reading them. A message about emergency treatment is just as likely to be addressed to “the GP of Mrs McTavish” as to “Dr Alan Finlay”, and even in the latter case, it's understood that the letter may well be handled by a secretary, partner or locum. A system that encrypted these messages would have to support the notions of roles and deputies.

In discussions of messaging, it's often presumed that the greatest security threat is that of someone eavesdropping or otherwise observing the document in transit. While this is doubtless a possibility, widely available mechanisms such as TLS[11] are usually adequate for this.

TLS, however, doesn't cover two cases: when the document is temporarily stored on a machine between the source and destination; and when it is at the destination waiting to be read. In either case, the need is the same as in storage systems: to prevent unauthorised access to sensitive documents while they reside on NHS systems. As with storage, cryptography is a way of addressing that.

2.2 Access controls and restrictions

With paper records, the security risk is somewhat limited. It is, of course, possible to access the records in one location, typically by social attack[1], but eavesdropping and large scale analysis of records is largely impractical.

Electronic records are different. Locating and copying a given record is usually much quicker than with paper, and can be done from anywhere, not just where the record resides. It also becomes feasible to proces records in large batches. This is useful for research and resource planning purposes, of course, but the capability is also open to abuse[2].

Wide-scale access to records raises other issues, too: for example, the risk of an error in a record being propagated and mistakenly acted upon is much greater when eyes other than the writer's may see it.

Here, we look at the attention electronic patient records and associated systems have received, first from legislators, then from the NHS and other clinical professional bodies. Finally, we look at two attempts by the security community to provide policy models that encapsulate these many rules.

2.2.1 Legislation

Given the importance of health records to individuals and to the nation as a whole, it's hardly surprising that many laws govern and control their use. An NHS guide to legal issues[22] lists fifteen sets of legislation as “the most relevant” to the use of electronic records. These range from the long-standing common law duty of confidentiality and the rules of evidence, to various Acts concerning public, private, and medical records, to Acts addressing the control of certain procedures and medical conditions.

One especially significant piece of legislation is the Data Protection Act of 1998 (DPA). The DPA places a number of controls and obligations on any body that collects personal information of private individuals. It requires such organisations to register the details and intended use of the data they collect with a central public registry, and to inform the subjects – those the data is about – before the collection takes place.

The DPA also gives private individuals the right to request a full copy of the data an organisation holds about them, and imposes limits on how much time an organisation can take and the charge it can impose. The Act allows a few exceptions including, in some circumstances, medical records.

There are just two reasons a doctor may withhold certain parts of a person's medical record [27]. The first is when releasing the information might lead to serious harm to the patient's physical or mental health, or to that of another person. The second is when releasing the record would reveal private information about other people or break a confidence with them. This might happen, say, with the birth and neonatal records of a young adult who had been adopted.

Another recent piece of legislation that has far-reaching implications is European Directive 95/46 on the processing of personal data, which the UK implemented in 1998. It affirms that patients have authority over their healthcare records, obliges record keepers to tell patients whenever access or changes to the records need to be made, and to obtain the explicit consent of the patients to do so.

It's important to bear in mind that the DPA and related legislation applies to all systems containing personal data. That includes information about staff – and researchers – as well as patients.

2.2.2 Guidelines and practices

In 1995, the UK government announced a programme to establish an NHS-wide network. One of its main objectives was to facilitate an electronic patient records system

While this promises to make it easier to provide fast and effective care to patients wherever they are in the country, it also increases the risk of inappropriate use or publication of patients' private data. A number of bodies have addressed this problem, and issued series of policies and guidelines for both clinical and other staff to follow when handling patient data.

2.2.2.1 The NHS Code of Practice

In 1996, the NHS issued guidelines for The Protection and Use of Patient Information. These were the first to cover the full implications of the new information and communications technologies, and have now been subsumed into the NHS Code of Practice for Confidentiality[26].

The Code of Practice describes a three step process: Protect - Inform – Provide Choice, reflecting both the duty of confidentiality and the provisions of the law. It also provides guides for conduct, practice, informing and obtaining consent from patients, and procedural tools such as decision flowcharts for use in a variety of circumstances. The guidlelines address healthcare use, non-healthcare medical use, such as research, public health or social services use, and non-medical uses, such as giving evidence in court or answering questions from the media.

2.2.2.2 The General Practices' guidelines

The two bodies representing general practitioners have also produced a set of guidelines for using electronic records in general practices[14]. The GPs' guidelines largely cover the same issues and principles as the NHS Code of Practice, though naturally with a focus on circumstances in primary care. However, they pay special attention to preserving the accuracy, integrity, and meaning of patients' records.

The guidelines consider records not just in their traditional role of communicating information to colleagues or to the future, but also in the new role as data for automatic processing. They address a number of practical issues: summarising and encoding information to facilitate automatic processing; updating or amending records, e.g., when a tentative diagnosis is changed; and transferring records between systems.

2.2.2.3 The Caldicott Committee Report

When the 1996 NHS guidelines were issued, the UK's Chief Medical Officer realised that the issues surrounding the other, non-healthcare uses of patient data needed to be addressed too. He commissioned Dame Fiona Caldicott to chair a committee to investigate and report on non-medical uses of PII. The committee identified and analysed over eighty data flows in applications such as medical research studies, work with public health authorities such as handling epidemics, and efforts with social services and other bodies to provide extended care to patients.

The committee's final report contained a set of six principles and fifteen recommendations for future practice. The principles alone make a nice summary:

> Formally justify every proposed use of PII;
>
> Transfer information only when absolutely necessary;
>
> Transfer only the minimum information needed for the task;
>
> Restrict access to the information to those with a need-to-know;
>
> Apprise everyone involved with the data of their responsibilities;
>
> Ensure everyone understands and complies with the applicable laws.

Some of the specific recommendations from the Caldicott report are also worth quoting:

> Protocols should be developed to protect the exchange of patient-identifiable information between NHS and non-NHS bodies.

The NHS number should replace other identifiers wherever practicable, taking account of the consequences of errors and particular requirements for other specific identifiers.

Strict protocols should define who is authorised to gain access to patient identity where the NHS number or other coded identifier is used.

Where particularly sensitive information is transferred, privacy enhancing technologies (e.g. encrypting identifiers or "patient identifying information") must be explored.

Where practicable, the internal structure and administration of databases holding patient-identifiable information should reflect the principles developed in this report.

2.2.3 Security Models

A security policy model is a formal, or at least rigorous, treatment of an information processing system. It encodes a number of properties and requirements of the system as a set of rules controlling which actors in the system may access which objects. The properties and requirements normally come from a careful analysis of the applications, their context, and the threats to the information and the associated costs.

Several security policy models are in common use. The Bell-La Padula model defines the multi-level security systems widely used in government and military applications. The Clark-Wilson integrity model is used for financial applications, and focusses on the integrity of data and accountability of its handlers.

More recently, a series of Role-Based Access Control (RBAC) models have appeared that address a wider range of organisations' needs. As the name suggests, a role-based model associates controls on information access with the roles people are performing, rather than with individuals.

As we have seen, access to patient records is governed by a complex set of rules. While Clark-Wilson and role-based models handle some aspects, they fall short of expressing all the nuances[1].

2.2.3.1 The BMA Security model

When the government announced its plans for the NHS-wide network and electronic patient record system, the British Medical Association (BMA) became very concerned about what provisions were being made to safeguard the confidentiality of patient information. They commissioned work to define a security policy that would handle the needs and interests of clinicians and patients alike. After considering many factors, including medical ethics, current and forthcoming legislation, and the threats and misuses of computerised medical records that had been observed to occur, the BMA Security Model was published[1][4].

One unique feature of the BMA model is that the subjects control access to information about them, not the keeper of the record. The patient is the authority over access to her records, in all but exceptional circumstances. In the model, each record carries an Access Control List (ACL) defining who may read or change the record and the individual responsible for maintaining the ACL itself. The patient must give explicit informed consent before any change to the ACL may be made.

A further feature of the BMA model is that patients must also be informed and give consent when someone wishes to use their records as part of a large selection. This reflects concern over confidentiality problems emerging from such analysis.

The difference in viewpoint to the NHS guidelines is subtle, but important. The NHS and Caldicott guidelines are NHS-controlled ("we hold this information about patients, and it's our duty to protect it and keep the patient in the loop"), whereas the BMA policy is patient-controlled ("the patient grants us access to these records on our advice for the benefit of the patient and society"). Both include provisions for exceptions and overrides, e.g. access in medical emergencies, or the rare cases where the public interest outweighs that of the patient.

2.2.3.2 The Tees Confidentiality Model

In 2002, the NHSIA commissioned a study of what patients understood about how their medical records were used, and how they would like that information to be managed in the future[28]. The purpose of the study was to provide a basis for further developments, including the NHS Code of Practice (see above), the electronic patient record system, and in particular, a consent management system.

The study found that, while people trust the NHS to take proper care of their records, they have little understanding of the uses to which the information is put. People are more concerned with who can see their information, and whether it is anonymous, than with what they do with it. Most people are comfortable with their GPs, hospital doctors, and emergency services to have access to their records, but feel that others, including specialised practitioners such as dentists and chiropodists, should only have access to the parts that are relevant to them, and that information used for purposes other than treatment should be anonymous.

Many felt that some parts of their records, particularly those concerning reproductive and mental health, are particularly sensitive, and want to have particular control over them. The study mooted the notion of a "virtual sealed envelope" a patient could use for such purposes.

The Tees Confidentiality Model (TCM)[21] is an authorisation model that embodies these findings. Its authors expect it be applied to the electronic patient records system under development, and to be a foundation of a future standard for medical confidentiality.

The TCM extends role-based access control with an explicit notion of Confidentiality Permissions, and with the concept of a collection that can be used to structure roles, individuals, and records and types of record. These make it possible to express not just general rules, but also the exceptions and overrides that are needed in complex situations.

2.3 The National Health Service as an environment

2.3.1 The staff

The NHS has over 1.2 million employees. A little over half of them are clinical staff, and roughly a third support them directly. Roughly one million staff potentially have access to medical records as part of their jobs.[16]

The staff are very mobile. Last year, for example, 200,000 staff changed jobs within the NHS, 75,000 joined the service, and 90,000 retired or left.

Like most people, NHS staff have time away from their jobs, for training, sickness, or holidays. The UK average is 9 days of sick leave per year, and the law prescribes at least 15 days of paid leave per year. On an average day, then, sixty to seventy thousand staff are on leave and a similar number are filling in for the absentees, either as temporary replacements or in addition to their normal duties.

Most staff have received training on confidentiality issues, the Code of Practice and Caldicott recommendations, and in our experience, take proper care of precautions such as locking workstations. However, while the staff are skilled and well qualified for their jobs, most have only application level knowledge about computing or IT.

So, for example, while clinical and administrative staff are well able to use email in the course of their jobs, the various error and security-related messages their systems and applications produce from time to time don't always mean a lot to them.

2.3.2 The workload

The NHS handles a million patients a day in primary care. In a year, hospitals and clinics handle nearly 13 million emergencies and minor injuries, and treat a similar number of outpatients; 11 million patients receive treatment as inpatients.

Each contact between a patient and a member of the clinical staff may eventually lead to an access or update to the patient's record; in the case of hospital stays, this could be several times a day. As well as the number of accesses to patient records, it's important to consider the number of authentication and authorisation operations involved.

2.3.3 The organisation

Changing organisational structures mean changes in reporting and responsibility. Authority and accountability are important concepts in any security-related design, not least in those that use cryptography, designing for change is essential.

From its start as fourteen regional hospital boards, the NHS organisation has been revised continually. In addition to changes brought about by growth and social change over the years, its role as a national organisation has subjected it to repeated reorganisations as governments and administrators embrace new political, economic, and social theories and policies. In the past twenty years, there have been twenty reorganisations[17]

Today, the NHS is organised as nearly four hundred hospital trusts and over three hundred primary care (general practices and clinics) trusts, supplying healthcare services to 28 so-called Strategic Health Authorities. More changes are certainly in the pipeline.

In a large, distributed organisation, it's inevitable that groups in different locations or functional units will face similar problems, but acquire different products and solutions to address them. These are unlikely to work together well. One of the functions of the NHS Information Authority (NHSIA) is to serve as a standards and selection body for the NHS for IT systems, to try to ensure that NHS systems interwork both internally and with external partners.

2.3.4 The systems

Basic operating software for desktop systems is standardised on one vendor, although systems are running versions up to eight years old, on the principle “if it's doing the job, leave it be.”

General practices are interesting cases. Each practice is responsible for selecting and operating its own systems. Administration is often outsourced, or performed by one of the staff as a second job; there is no central IT administration or support. Some systems, such as the mail exchange serving GPs in a region, is operated by the health authority.

The practices run specialised applications that handle the routine chores: managing patients and appointments, maintaining patients' records, producing standard letters for referrals and other purposes. These are designed to cope with current working practices, it's sometimes not clear how to extend them to cope with new ones.

We found, for example, that the system used by one practice could automatically generate a referral letter from a doctor's notes and a template file. While it could easily print the letter, we found no way to email it, short of saving it as a file and sending it as an attachment. To the secretary who normally did the work,this was a major departure from what she was used to.

3 Cryptography for information systems

In this section, we examine two public key cryptography systems and how they stand up to the challenges the NHS poses.

We look first at conventional public key cryptography, for which there is already a body of accumulated experience in trying to apply it to organisational use. Then we consider how identity based PKC, which promises to solve or avoid several of the problems PKC encounters.

3.1 Public Key Cryptography and PKI

In conventional Public Key Cryptography, someone wishing to receive and decode encrypted messages must first register with a trusted third party knowns as a Certificate Authority (CA). After authenticating the applicant, the CA issues them with a pair of keys, one to be kept private, the other to be made public. The CA also issues a certificate, which confirms that the public key is associated with that person, or at least, with an identity.

To encrypt a message for someone, a sender first has to get the recipient's public key; this might be published, on a web site, say, or the recipient might supply it on request. The sender then has to check the public key is genuine, by checking the certificate, and valid, by looking for the certificate on a revocation list the CA issues periodically. Once satisfied, the sender encrypts the message using the public key, and sends the encrypted text to the recipient, who is then able to use the private key to decrypt it.

A Public Key Infrastructure (PKI) is a system intended to provide the key lookup and verification services to a large group of people. X.509 is a widely-used specification for digital certificates. S/MIME defines a way of using these to encrypt email. Unfortunately, both X.509 and S/MIME admit a number of interpretations; this has led to a number of incompatibilities between different vendors' products.

3.1.1 Attribute certificates

Attribute certificates (ACs) [13][30] are a recent addition to the X.509 digital certificate specification. They are intended to support attributes beyond identity, including roles and capabilities, and are designed to be more flexible and short-lived than identity key certificates. ACs are designed to be used in conjunction with directories and similar systems: to check if someone has a particular capability, one checks if the corresponding attribute certificate appears in their directory entry.

ACs allow delegation of authority. However, the processing for verifying delegations can become expensive if the role hierarchies or chains of command or delegation are even moderately complex[19].

Some experimental systems are under development. The PERMIS project[8] is an authorisation system that uses attribute certificates, and which has been applied successfully in a number of trials and contexts, including some medical applications in parts of the NHS.

3.1.2 How does PKC fare in the NHS environment?

In 2001, the NHS Information Authority launched a project to build a PKI for the NHS[24] intended to be fully deployed and operational in mid-2002. It was to support the usual services including authentication and secure messaging. Within the year, it was clear that the practical and logistical problems were much larger than had been thought at first[25] and the schedule was reviewed and lengthened.

Implementing a PKI is widely accepted to be a complicated job requiring careful analysis and planning, and considerable time and expense. Even in organisations much smaller than the NHS, the take-up of PKI solutions has been slow; a recent survey[12] indicated a number of reasons, including poor expected return on investment, lack of application support, and the difficulties associated with deployment and use. These largely echo the findings of the NHSIA effort.

3.1.2.1 Scale issues

The usual problem with conventional PKC is how to manage keys. First, there's the problem of managing the natural mobility of the staff. Remember that roughly 200,000 staff changed jobs and a further 90,000 left the NHS in 2002. The usual lifetime for a certificate is one year, so a quick calculation suggests that a central NHS PKI would have a certificate revocation list containing 45,000 entries on average. That is just to cover people who have left the NHS, not temporary changes such as sickness cover.

In [3], Anderson discusses a number of attempts to manage trust centrally in banking and healthcare settings that proved at best difficult, even though the organisations were small relative to the NHS.

Another problem is an apparent mismatch between the granularity of the keys and the needs of the application. PKIs normally associate keys with individual users, whereas the need is to address and deliver messages to roles or groups[7]. Obviously, it is possible to assign a key and a certificate to a group, but that gives rise to another problem: as a group normally has several members, access to the group key must then be shared among several people. What happens when one of them leaves the group? How does the system audit individual use of the shared keys?

Attribute certificates have been suggested as a potential solution: attributes could grant individuals access to shared keys, or even to the resources themselves. At a functional level, they appear to work; it isn't clear, however, how well ACs will perform in such a large and complex organisations.

3.1.2.2 Temporal issues

Managing keys over long periods of time also seems difficult. In most PKIs, the standard lifetime of a public key certificate is one year. While it isn't strictly necessary to regenerate the key when renewing the certificate, this nevertheless seems fairly usual. Each user gradually accumulates a set of keys which, even though expired, have to be kept in order to access older documents.

Thus, a large archive has to store and manage not only the data, but all the keys needed to decrypt the records as well. It would, of course, be feasible to re-encrypt the data under a common archive key. Such a scheme, though, would have to be conform with the many confidentiality rules and guidelines. Either way, it's a difficult problem to solve.

3.1.3 In summary – public key cryptography

It seems the perennial problem with conventional PKI and PKC systems is the management of keys, particularly with large systems and over extended periods of time. This has been the case for several years now, and little progress seems to have been made.

Many problems appear to be interworking issues, arising from different interpretations of the X.509 specification, among others. Avoiding these problems without becoming locked in to a single supplier (an issue for a public service) is a great challenge. It's hard to see how it will be resolved.

3.2 Identity and Identifier Based Encryption

In the mid-eighties, seeing that wide use of cryptography was being held up by the problems in distributing and managing public keys, Shamir posed a problem: to create a public key cryptosystem that needed no PKI[34]. In 2001, two groups of researchers[5][10] independently produced practical solutions to the problem. Since then, the two groups have embraced the same underlying mathematics – pairings – but have pusued somewhat different approaches to using it.

In previous public key encryption schemes, when one person wishes to encrypt a message for another's eyes only, either both parties have to agree a shared key to use in advance, or the sender has to obtain the recipient's public (encryption) key, check its validity, then use it to encrypt the message.

In an identity based schemes, the encrypter chooses some string of characters. The chosen string, whatever it says, is the encryption key, and obtains a set of encryption parameters from a trusted third party: a Key Generation Centre (KGC). The encryption function takes the string and the KGC's parameters, and emits the encrypted text.

To decrypt a message, the receiver also needs to know the encryption key string, and the KGC used by the encrypter. The would-be decrypter then needs to persuade the KGC to supply the decryption key corresponding to the key string. It's a point of trust in an IBE system that the KGC will take proper steps to ensure it doesn't give the key to an imposter.

Note that the KGC can generate a the decryption key corresponding to a given encryption key string at any time. This has important consequences. It means that key escrow is a basic property of IBE. Whether this is desirable depends on the application and on the users; clearly, they will have to trust that whoever operates the KGC does not misuse this power.

It also means the sender needs no prior contact with the recipient. Provided everyone using a KGC observes the same convention for choosing the encryption key string, a sender can encrypt messages for someone before the intended recipient has obtained a decryption key—or even an email account[31].

As we have described it thus far, once I have the key corresponding to my email address, I can use it to decrypt all messages encrypted and sent to it—and so could anyone else. Clearly, it's important for people to keep their decryption keys safe and secret. In practical terms, one might prefer to limit the damage losing a key could do, say, by limiting its lifetime.

This turns out to be easy to do. Since the string used to encrypt the message can be anything, the users of a KGC simply have to adopt a different convention, e.g., to catenate the email address with some other value, such as a date, or even a unique message number.[5]

If we explore further what conventions might be useful for choosing encryption keys, we begin to see a variety of possibilities for creating decryption keys that someone could dedicate to special circumstances or applications.

If we shift the focus of the encryption string from the recipient to the message, the possibilities become even more interesting: we can use the encryption string to describe the circumstances, or policy, under which someone may be allowed to decrypt the message.

The approach was first described in detail by [20], and is the the one we have been following in our work at HP Labs. This approach is sometimes referred to as Ident*ifier*-based encryption, or Identifier-based public key cryptography (ID-PKC).

In this approach, we refer to the TTP as the Trust Authority (TA). It serves two roles. As before, the TA is a KGC: it generates the decryption key corresponding to a given string. But the TA also acts as a *reference monitor*[18], a single point that makes the access control decision for an object. In this capacity, the TA can invoke or implement practically any access control scheme or policy.

The "identifier" is thus a message from the encrypter to the TA that describes explicit details of the access policy the TA, as reference monitor, should enforce. The encrypter trusts the TA to carry out that policy, but also understands that the TA interprets it in a wider context of rules, laws, and policies that varies over time. These might include standing rules for escrow, delegation and revocation of authority, or rules that reflect new laws. Thus, while the would-be decrypter may match the description given by the encryption string, the TA may decide not to hand over the decryption key.

3.2.0.1 Work flow and Split Authority

One of the appealing things about ID-PKC is that systems can be built that use more than one trust authority. This creates some useful possibilities[20].

The first is cryptographic support for enforcing work flow, that is, in order to receive the full key for a message, a would-be decrypter has to follow some series of steps described by the encryption string: to visit a number of parties to verify credentials or to confirm that certain important steps have been carried out.

The second splits the authority between several TAs. On a request for the decryption key, each TA makes a decision based on its own standing policies, and generates a partial decryption key; the partial keys from all the TAs are needed to decrypt the message. This offers a way round the built-in key escrow of a single TA: each TA in the group effectively has a veto.

3.2.1 How does IBE fare in the NHS environment?

3.2.1.1 Staff issues

IBE copes naturally with role changes and the comings and goings of staff. The authorisation check can be as simple as checking whether someone has a given role listed in a directory or database. At first sight, this may seem like cheating, in that it's pushing the load of making the decision onto some other system. We feel, though, that the operations involved are typically simple and in many cases, are likely to be performed anyway.

IBE seems to be straightforward for people to use. Finding the encryption key can be as easy as finding the email address of the recipient[35], and even though the encryption key may contain a much more complicated expression of policy, it seems that for most applications, this can be generated automatically, with the configuration left to the IT or information governance staff.

3.2.1.2 Scale issues

Our prototype messaging system implemented the trust authority as a secure web service. The cryptographic operations are comparable in computational effort to those used in current PKI systems. Web servers and services capable of handling thousands to millions of requests per day are increasingly common, as are the techniques and tools used to build them. Naturally, much depends on design choices for the system, particularly whether to centralise or distribute the TA function. Distributed TAs seem preferable for performance, robustness, and security reasons.

In most of the literature to date, the choice of trusted third party (KGC or TA) rests with the person encrypting the message. This may well be the case for general-purpose messaging systems. In organisational use, such as in the NHS, a sender is likely to use the TA most closely associated with the intended recipient. Since the TA is supposed to verify the identity and credentials of the decrypter, in seems natural to site the TA in an environment where it can make those checks. TA's are likely to be established on organisational lines, just as mail and network security domains are currently.

As with many systems, there are interesting problems in the details.

First, how does the encrypter of a message find the TA appropriate for the destination of the message and obtain its parameters? The obvious answer, and it seems a reasonable one, is to hold this information in a NHS-wide directory (one is currently being developed), along with the addresses of roles and groups. Other network information services could be used: the domain name service (DNS) caries information about mail servers, and could easily be extended to do the same for TAs[35],

Choosing the the language of the encryption key strings is a more complicated issue. Because the string is a message across space and time to the Trust Authority, it's important that both parties can parse it and take the same meaning from it. What happens when one party uses “secretary” to describe what the other calls “assistant”, or when the policy requests approval from an organisation that disappeared two NHS reforms earlier?

In a somewhat more general form, this is the “service discovery problem” of distributed systems. In practice, we would expect a body such as the NHSIA to issue and maintain standards for the vocabulary, syntax, and semantics of these messages.

3.2.1.3 Temporal issues

The language problem is significant over time, too. provided we are able to establish who the contemporary equivalents are for the principals named in the original encryption string, it shouldn't be a problem to determine whether to grant access.

One thing that could be a special concern is the sensitivity of IBE or ID-PKC to a compromise of the system secret (this is a parameter, known only to the TA, from which the public encryption parameters and all decryption keys are derived). If this is exposed, all messages encrypted using the TA are at risk of exposure.

This is a problem for any cryptographic system expected to operate over a long period of time. It is usually straightforward, if tiresome, to issue a new key or switch to a new system secret. The records, potentially numbering in the millions, will remain encrypted under the old key. How to reinstate their security quickly and efficiently is an interesting open problem for both cryptographers and system designers.

3.2.1.4 Regulatory issues

The TA evaluates the message in the encryption key on each request for the decryption key. This allows the TA, as reference monitor, to take account of all the factors that apply at the time. In principle, it is straight-forward to comply with new bodies of regulations as they appear.

Audit is also a significant part of current regulatory requirements. Again, this should be straight-forward to build in a TA.

3.2.1.5 Practical issues

The major practical issues in deploying an ID-PKC system that we haven't already addressed is that of implementing the reference monitor. How does the TA calculate whether to grant someone the decryption key for a given message or not?

In practice, the TA would not actually handle the decision itself. Instead, the TA would be designed to call out to a decision-making component of whatever policy or authorisation management system is appropriate, e.g., one implementing the BMA security policy model or the TCM.. This even allows a degree of future-proofing, in that it should be possible – with careful design – to replace the decision module with newer ones as needs dictate.

3.2.2 In summary – identifier based cryptography

ID-PKC has the flexibility to support the practical needs of NHS workers. It naturally supports addressing to roles and groups as well as to individuals and has none of the key distribution or management problems traditionally associated with secure email. It's straightforward to see how to incorporate it in systems that need to handle eternally shifting role to person mappings. The mechanisms needed to handle issues of scale and robustness are, for the most part, similar to those already in use for handling busy web sites on the Internet.

In storage applications, ID-PKC allows a close binding of a record and the policy that controls access to it, a feature that seems desirable in the BMA and TCM policy models. Yet by referring the access control decision to an external policy engine, it gives the flexibility to cope the changes that may occur over time, for example, in legislation or in the organisation's structure.

ID-PKC's support for split authority is clearly useful. We have seen that different bodies in the NHS perceive different potential uses for patient data: the clinical staff to improve care, and administrators to help plan services and make difficult decisions over future investments, among other uses. Such groups might be uncomfortable with one having control over a single TA. Split authority would allow each to maintain a TA, with different standing policies and independent auditing; thus, each group could ensure that its interests were at least being considered.

The other interesting feature of ID-PKC, work flow support, may be difficult to exploit, at least for applications involving patient records. This is because of the long life of the data. Work flow descriptions inevitably reflect assumptions about a particular application and organisation; after even a few years, these might not hold true. As we have seen, even expressing access control policies that will apply for a number of years is an interesting problem.

Indeed, the passage of time provides some interesting questions for ID-PKC, as for any encryption scheme: how to keep system keys and secrets safe for long periods, and how to manage the transition of systems and large numbers of documents to new keys, should it become necessary. Some recent research [6][15] looks promising; there is still plenty left to do

4 Conclusions

The NHS is a challenging environment for public key cryptosystems: its size, the mobility of its staff, the eternally shifting organisation, the sensitivity and value of the data it handles, and the complex rules that govern it all. For someone looking to apply cryptography to practical problems, it offers much to study and learn from.

We set out to qualify our instincts that identifier-based public key encryption could cope with the practical issues posed by the NHS environment. On the whole, we feel we have done this; for most of the problems posed, we can sketch likely solutions, even though actual implementations may be large tasks.

A particularly interesting set of research problems comes from the long life of patient records. Can we design practical, robust, elegant cryptosystems to protect large populations of data across decades? The challenge is certainly there.

5 Acknowledgments

We would like to express our thanks to all those whose suggestions and efforts contributed to this work. In particular, thanks to Nigel Harvey, Nigel Hudson, Aled Edwards, Kenny Paterson, and Jonathan Griffin for their helpful comments on early drafts of this paper.

References

1: Anderson, R J. *A Security Policy Model for Clinical Information Systems*, 1996
2: Anderson, R J. *Patient Confidentiality - At Risk from NHS Wide Networking*, 1996
3: Anderson, R J. *Problems with the NHS Cryptography Strategy*, 1996
4: Anderson, R J. *Security in Clinical Information Systems*, 1996,
5: Boneh, D; Franklin, M. *Identity based encryption from the Weil pairing*, , 2003,
6: Canetti, R; Halevi, S; Katz, J. *A Forward-secure Public Encryption Scheme*, *Eurocrypt*

2003, 2003,
7: Casassa Mont, M; Bramhall, P; Dalton, C R; Harrison, K. *A Flexible Role-based Secure Messaging Service: Exploiting IBE Technology in a Health Care Trial*, 2003,
8: Chadwick, D W; Otenko, A; Ball, E. *Implementing Role Based Access Controls Using X.509 Attribute Certificates: the PERMIS Privilege Management Infrastructure*, 2002
9: Chen, L; Harrison, K; Soldera, D; Smart, N P. *Applications of Multiple Trust Authorities in Pairing Based Cryptosystems*, 2002
10: Cocks, C. *An Identity based encryption scheme based on Quadratic Residues*, 2001
11: Dierks, T; Allen, C. *The TLS protocol* , 1999
12: Doyle, P ; Hanna, S. *Analysis of June 2003 Survey on Obstacles to PKI Deployment and Usage* , 2003
13: Farrell, S; Housley, R. *RFC 3281 - An Internet Attribute Certificate Profile for Authorization* , 2002
14: General Practitioners' Committee and Royal College of General Practitioners. *Good Practice Guidelines for General Practice Electronic Patient Records* , 2000
15: Gentry, C; Silverberg, A. *Hierarchical identity-based cryptography*, *AsiaCrypt 2002, LNCS 2501*, 2002,
16: Government Statistical Service. *Staff in the NHS 2002*, 2002
17: Institute for Public Policy Review. *Press Release 236 - NHS: 20 re-organisations in 20 years* , http://www.ippr.org.uk/press/index.php?release=236
18: Irvine, C E. *The Reference Monitor Concept as a Unifying Principle in Computer Security Education*, 1999
19: Knight, S; Grandy, C. *Scalability Issues in PMI Delegation*, *1st annual PKI Research Workshop*, 2002,
20: Levy, I, Identifier based PKC - Potential Applications, 2002
21: Longstaff, J J; Lockyer, M;Nicholas, J. *The Tees Confidentiality Model: an authorisation model for identitiesand roles*, *SACMAT'03*, 2003,
22: NHS. *Using Electronic Patient Records in Hospitals: Legal Requirements and Good Practice*, 1998
23: NHS. *For The Record: NHS Retention & Disposal Schedule* , 1999
24: NHS. *Strategy for cyrptographic support services in the NHS* , 2001
25: NHS. *NHS Cryptographic Support Services Update - March 2002* , 2002
26: NHS. *Confidentiality: NHS Code of Practice* , 2003
27: NHS. *Guidance for Access to Health Records Requests under the DPA 1998* , 2003
28: NHS Information Authority, Consumers Association. *Share with Care - People's Views on Consent and Confidentiality of Patient* , http://www.nhsia.nhs.uk/confidentiality/pages/docs/swc.pdf
29: NHS, North-west London Hospitals Trust. *Information for GPs* , http://www.nwlh.nhs.uk/sitecontent/gps.html
30: Nykaenen, T. *Attribute Certificates in X.509*, ,
31: Outscheme Inc. *Mailinator* , http://www.mailinator.com/
32: Paterson, K G. *Cryptography from pairings: a snapshot of current research*, 2002
33: Rivett, G. *NHS History* , http://www.nhshistory.net/
34: Shamir, A. *Identity-based cryptosystems and signature schemes*, 1984
35: Smetters, D K; Durfee, G. *Domain-Based Administration of Identity-based Cryptosystemsfor Secure Email and IPSEC*, *12th USENIX Security Symposium*, 2003,