A Simulation Model for Evaluating National Patient Record Networks in South Africa

by

Kieran Sharpey-Schafer

Supervised by: Dr Hussein Suleman

Thesis presented for the Degree of Master of Science
In the Department of Computer Science
University of Cape Town

November 2012

I know the meaning of plagiarism and declare that all of the work in the thesis, save for that which is properly acknowledged, is my own.

Kieran Sharpey-Schafer

Acknowledgements

This research would not have been possible without the input of a few important people. My sincere and appreciative thanks go to:

My supervisor, Hussein Suleman – a consummate professional and an expert in.... well everything. His reliability for constructive feedback and ideas was invaluable, and a great learning experience in itself.

My family and Michelle, for giving me the opportunity, love and support to make this happen.

Fellow MSc students, past and present, for their insight and encouragement in stormy weather. A particular reference is saved, footnoted and reformatted for the help of comrades Alex Salo and Sarah Brown.

Andreas Varga and the thriving OMNeT++ community – a great example of successful open source software.

Lastly, my thanks go to Cell-Life for their continued support of my studies and introducing me to the fascinating world of health informatics.

Dedication

This work is dedicated to all those working on improving healthcare in the various challenging contexts around the globe.

Abstract

Access to an accurate medical history is criticalto the quality of patient care. Lack of knowledge of a patient's medical history can cause complications in treatment and delays in diagnosis. Nationally available patient medical histories can help prevent this, as healthcare professionals can share and maintain a single record. Referred to as a National Patient Record (NPR), electronic networks providing this service are being implemented by many countries globally. The approaches however, have many different architectural designs and few have been tested in or designed for the unreliable infrastructure of developing countries.

This research proposes a simulation model for evaluating National Patient Records in different contexts. The model can be used to represent the algorithms of various NPR network designs, and then simulate them in the particular context of South Africa. The model was developed using the simulation study methodology – collecting data on the real world systems to build a conceptual model, programming the simulation model and then improving these through iterative validation and testing.

The NPR model created was shown to reasonably represent the behaviour of two NPR algorithms as well as the South African population and local connectivity challenges. The NPR model was then successfully used in a case study to evaluate two particular NPR architectures: the model of the UK centralised NPR design and the US peer-to-peer NPR design. The two models were tested at different scales of load, where the US model was observed to handle growth and downtime challenges more consistently than the UK Model. In the case study, the NPR model demonstrated how it can be used as an additional evaluation tool for assessing the performance of potential NPR systems for South Africa.

This study has shown that modelling and simulation is a feasible approach for evaluating NPR solutions in the developing context. The model can represent different network models, patient types and performance metrics to aid in the evaluation of NPR solutions. Using the current model, more case studies can be investigated for various public health issues – such as the impact of disease or regional services planning.

Table of Contents

List of Figures

List of Tables

Glossary of Terms

DoH	Department of Health
EPR	Electronic Patient Record
GP	General Practitioner
MPI	Master Patient Index
NHI	National Health Index- used in New Zealand to register all patients
NHIN	National Health Information Network– the US's program for nationally available Electronic Patient Records
NHS	National Health Service – the UK's public health service
NPfIT	National Program for IT – the UK's program for nationally available Electronic Patient Records
NPR	National Patient Record
Obsolescence	The number of changes to the most current record version that have not yet reached the local copy
OMNeT++	An open source discrete event simulation framework written in C++
PHC	Primary Healthcare Clinic
RNG	Random Number Generator
RTT	Round Trip Time
SA	Sub-Saharan Africa
UK	United Kingdom
US	United States of America
IV&V	Independent Verification & Validation
VVT	Validation, Verification and Testing
SME	Subject Matter Expert
UT	Unit Test

Chapter

1 Introduction

Health workers rely on medical histories to assist in their diagnosis and care of patients. An ideal medical history would be one that kept record of the patient's entire interaction with all healthcare services in their lifetime, that was accessible nationwide. While some facilities do keep medical histories in an Electronic Patient Record (EPR), the majority of health care institutions have no method of sharing them. This inability to share information becomes a problem for two reasons: firstly, records cannot be easily accessed when a patient moves between two institutions; and, secondly, it is difficult for government to be able to compile accurate statistics on health in South Africa. The South African government has indicated that they have begun planning to introduce a *National Health Information System*for sharing and collecting patient records on a national scale(South African Department of Health, 2001).

Most public health facilities still rely on paper-based systems that record a patient's history of care, and the memory of the patient to recount their various encounters. Those public facilities that do use electronic systems to record patient histories are seldom able to share these with other facilities – mainly due to the variety of different systems being used and the lack of any incentives for interoperability.

1.1 The Vision of the Electronic Patient Record

When healthcare professionals have access to a single, unified copy of the patient's medical history, they are able to make better informed decisions on the needs of the patient. Furthermore, the practicalities of patient management are also improved by digitising patient histories – as shown by a Medical Records Institute survey (Medical Records Institute, 2002) that showed doctors using paper-based systems spend 38% of their time doing 'data management', which can be significantly reduced when the records are made electronic. These are the main arguments for the maintenance of a patient's electronic medical history, which is more commonly known as an Electronic Patient Record (EPR).

Implementing collections of all EPRs on a national scale – a collective National Patient Record (NPR) – has even further benefits. National monitoring of disease, epidemics and general health can be greatly improved in terms of accuracy, coverage and response time. Analysis by region or population demographic becomes possible and can potentially lead to targeted public health initiatives which would previously have been neglected.

Global efforts to implement NPR systems have been on the increase yearly since the early 1990s. Denmark, New Zealand and Canada already have whole or partial implementations in use, and no less than 20 other countries are in the process of building or designing similar systems (Accenture, 2006).

1.2 The Diverse Designs of National Patient Records

Despite the broad international investment in EPRs and interoperability standards, the approaches used have varied considerably between countries. This can be clearly seen in the designs of two of the largest NPR projects – those of the United Kingdom and the United States.

The UK has attempted to support its entire National Health Service (NHS) with several applications that run on top of a single dedicated national network for healthcare systems. Known as 'The Spine', it is planned to be a nationally available EPR repository that local applications can use to retrieve the latest information on a particular patient. The program as a whole, known as the National Programme for IT (NPfIT), has a fundamentally *centralised* network architecture where the central server keeps a unified copy of a patient's EPR(Ringholm, 2007).

In direct contrast to the UK's approach, the United States has taken a *de-centralised* approach to sharing patient records. The plans for the National Health Information Network (NHIN) include regional gateways that will allow different local systems to share records adhering to a national metadata standard. The regional gateways are linked with other regions in apeer-to-peer network to allow nationwide sharing and lookup of patient records(Gartner, 2007).

1.3 The Unique Challenges of the South African Context

A nationwide system faces many implementation challenges in any country, regardless of the architectural design. These come in the form of costs, standards, stakeholders, politics, as well as the inherent barriers to change that the complexities of health care services can create.

Implementing an NPR in South Africa has many additional environmental barriers and constraints that make the operational context quite different to that of the US and UK. Basic challenges, such as weaknesses in infrastructure and human resource capacity, are combined with more complex issues such as different disease burdens and different patterns of healthcare usage.

Before an NPR can be implemented in South Africa, potential solutions, and how well they can perform in its unique context, need to be considered.

1.4 Simulating National Patient Record Systems for South Africa

Simulation has been widely used to analyse various network problems, from the feasibility of new packet level protocols to the load capacity of cell phone towers. In healthcare it has been used to model the processing queues in health facilities (Swisher, 1997) as well as the progression of disease outbreaks (Canner, 1992). However it has not yet to been considered in relation to this problem – evaluating appropriateness of a particular network solution.

Simulation could potentially help when evaluating the relative qualities of the different NPR systems for use in South Africa. If each of the candidate systems from across the world - or new additional designs - could be adequately represented in a simulation, they could potentially be studied for their efficacy and efficiency in the specified South African context.

The alternative approaches to evaluation by simulation are traditionally theoretical calculations or the building of real world prototype systems. For investigating the performance of different candidate NPR solutions in South Africa, building even a reduced prototype of each system would require vast investment and is thus somewhat infeasible. Alternatively few mathematical models can represent all the possible states of such complex distributed systems and accurately predict system behaviour. Thus, simulation could provide a more feasible method to analyse the proposed solutions, with controlled costs. As such, this research addresses the following research question:

"Is simulation a feasible technique to evaluate designs for a National Patient Record in South Africa?"

1.5 Research Outcomes

In answering this research question, the following sub-questions were also considered:

1. How can the various potential solutions for NPR systems be represented in a simulation model?
2. How can the various contextual constraints of a developing country like South Africa be represented in a simulation model?
3. How can a potential solution's performance in the South African context be evaluated?

Furthermore, the following results and artefacts were also obtained:

a) A programmed NPR model simulation tool.
b) A case study evaluating the scalability of a centralised NPR and a de-centralised NPR in South Africa.

These outputs collectively can provide the required information for assessing the feasibility of using simulation to assist with the evaluation of NPR networks.

1.6 Methodology

As shown in Table 1, the research began with gathering data on contemporary NPR efforts as well as the operational challenges faced by networks in the developing country context. This data then was used to inform the design of a conceptual model that can represent the fundamental components and behaviours of the networks. To test and evaluate the conceptual model, a number of solutions were designed and then represented in a programmatic model. Network and performance metrics are added to the programmatic model so that its behaviour can be quantitatively measured.

Once the accuracy of the programmed model was established, it then was used in a simulation study to compare the various network solutions' performance characteristics.

Methodology Stages
1. Research NPR solutions
2. Extract common requirements
3. Collect data on contexts
4. Build network models
5. Build context models
6. Build metrics
7. Iterate design steps 1-6
8. Validate Model
9. Simulation Study

Table 1: Methodology Stages

1.7 Scope and Limitations

Any network solution can be evaluated on many levels – from the lowest level bit transfer to the usability of the application layer. This research, however, focuses on the network architecture and message passing algorithms, and their impact on the performance of comparative solutions in the same context. Furthermore, this research focuses on the simulation study process of design, implementation and evaluation, and its feasibility in the new problem domain of National Patient Records. The research did not include scope to build an exhaustive conceptual model or industry ready software.

Additionally, this research was confined to the capacity and limitations of the OMNeT++ simulation tool, which was used for network modelling and simulation.

1.8 Thesis Structure

Background theory and literature on this problem are presented in Chapter 2. The chapter contains an explanation of contemporary NPR efforts, and their common set of core requirements, and is accompanied by the documented challenges of the South African context. How networks are evaluated traditionally and precedents for simulation studies also are explored. Chapter 3 outlines the design of the comprehensive NPR model that represents the network solution, local population and contextual challenges. Chapter 4 documents an investigation into the completeness and correctness of the programmed model in order to understand its accuracy, validity and limitations, as well as establish with what confidence the model can be used to evaluate NPR solutions. To illustrate the model's potential use in research, Chapter 5 shows how a simulation study was conducted using the model to compare two network solutions operating under the same conditions. Conclusions are presented in Chapter 6 by reflecting on the findings, outcomes and limitations of the research, and suggesting future areas of research that would be valuable extensions to this work.

Chapter

2 Background

There are many possible solutions for implementing a National Patient Record system in South Africa. All solutions aim to provide the generic benefits of health information systems, but their technological designs can be quite different. This chapter provides contextual information on the problem domain of information management in healthcare and the different approaches and designs that can be considered. The current challenges of the South African context that can impact the NPR design are discussed, as well as some background on simulation as an evaluation technique.

2.1 Information Management in Health Care

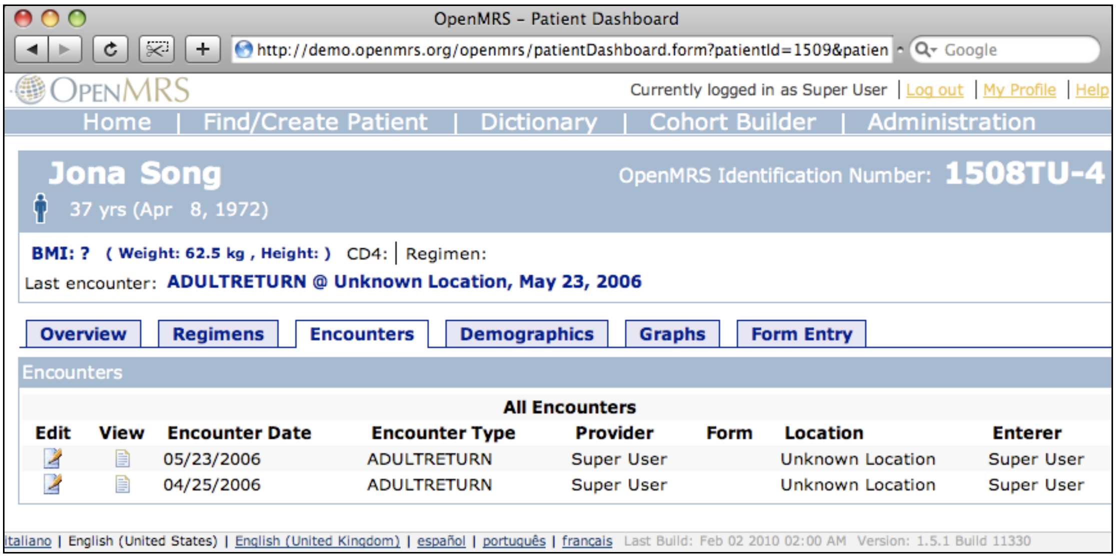
Patient records have been in use since formal medicine began. A traditional paper-based medical record has two basic functions: accumulating data on a patient's history and the co-ordination of procedures for the patient's care (Berg, 2004). Examples of patient data includes demographics like date of birth or medical information such as a history of illnesses. Co-ordination of procedures involves scheduling appointments for procedures or consultations or prescribing of drugs. Both help track a patient's progress through care, often known as a 'medical history'.

Similarly, the data collected in a patient's record can be used for different purposes, and Berg (Berg, 2004) distinguishes between them as follows:

- *Primary purposes* are those that use the information to assist health care professionals in the care of the patient. This might include data on lab results, patient demographics or a history of previous operations.
- *Secondary purposes* include all those that are one step removed from the primary care process, such as administration, research, accountability, management purposes, and so forth.

2.1.1 Electronic patient records

An Electronic Patient Record (EPR) fundamentally is an electronic version of the traditional paper-based medical record: capturing data and allowing co-ordination of procedures over a patient's time in care. However having the patient's record in an electronic format can provide many additional benefits that can potentially increase the quality of care. Some such improvements include easier sharing of information, broader standardisation, automated analysis and decision support (Berg, 2004). Secondary uses of data also become more efficient as a result of electronic records, such as monitoring the statistics of usage and illnesses at facilities.



Jona Song
37 yrs (Apr 8, 1972)
OpenMRS Identification Number: **1508TU-4**

BMI: ? (Weight: 62.5 kg , Height:) CD4: | Regimen:
Last encounter: **ADULTRETURN @ Unknown Location, May 23, 2006**

All Encounters

Edit	View	Encounter Date	Encounter Type	Provider	Form	Location	Enterer
		05/23/2006	ADULTRETURN	Super User		Unknown Location	Super User
		04/25/2006	ADULTRETURN	Super User		Unknown Location	Super User

Figure 1: OpenMRS - An Example of an Electronic Patient Record System

In Figure 1 above is an example of a Web-based EPR system. Users can access the system, from authorised Internet-connected computers, to find patients and their detailed medical histories. In the example above, the patient's demographic details are shown along with a list of their recent visits (here called "encounters") at various health facilities. EPR systems such as this often are used only within a specific facility, and seldom across facilities at the application level.

2.1.2 National Patient Records

A National Patient Record (NPR) is a national repository of EPRs, with one unique record per member of the population, which are available to all healthcare facilities. Rather than being based at the application level, as in the example above, NPR systems often work as a 'broker" of EPR records to various subscribing applications at the different facilities. Thus the NPR network focuses on maintaining records of the individual EPRs across all facilities, and other applications provide services on top of this.

2.1.3 Stakeholders

There are three main stakeholders involved in health care in any country, namely: the population (as patients), health care professionals and government. EPRs, and a national EPR in particular, can be of potential benefit to all three. A study of South African patients performed by Accenture in 2006 (Accenture, 2006) showed that almost a third of all participants used multiple health care providers and half said that they answered the same questions on each visit to a new practitioner. In this instance, a comprehensive NPR would allow healthcare workers to see a patient's historyand potentially reduce the number of questions repeated. This could reduce visitation time (and in turn reduce waiting times) and potentially improve care as a record completed by another health professional would presumably be more accurate than that recounted by a patient(Accenture, 2006). Health care workers themselves would benefit from an NPR – similarly benefiting from medical histories being quicker to obtain and more comprehensive, giving them more time and information with which to perform their responsibilities. The management of information, particularly documentation and data retrieval, is actually a major burden for most healthcare workers. A survey of 2096 medical residents from 263 medical centres in 47 American states, conducted by the Medical Records Institute(Medical Records Institute, 2002), showed that, over a 3 year period, 38% of their time was taken up by ‘data management’ (see Figure 2).

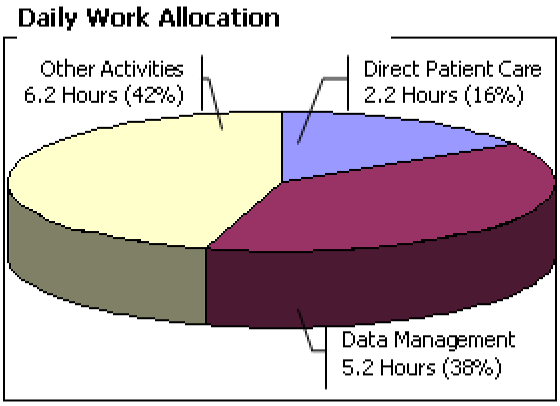


Figure 2: US Resident doctor average daily work allocation

This data management involves collecting and recording patient information, as well as periodically updating it based on clinical care(Medical Records Institute, 2002). Examples of this would include finding and gathering test results, laboratory reports and medication lists. Furthermore, the residents often spent time converting data into more mobile and usable formats. A comprehensive patient record system could possibly reduce time spent on retrieval of data.

Government also benefits from the implementation of a NPR. Without a national network of access to health information, any data that is captured, whether in electronic or paper format, only is accessible at the health care facility in which it was created. These ‘silos of clinical information’ are difficult to survey and thus make it difficult for the government to monitor health care delivery as well as potential disease management (Jones, 2004).

2.1.4 System Requirements

In terms of the requirements of a system it is important to distinguish between the *functional* and *non-functional* requirements. According to Puigjaner (Puigjaner, 2003), a functional requirement is what the user *needs the system to do* and a non-functional requirement is a description of a preferred *environmental constraint* under which the system needs to run. In an NPR system, an example of a functional requirement could be ‘requesting a patient history’ while a non-functional requirement (for health care workers) of *performance* might be that the requests are met with a rapid and timely response or, even more specifically that all results are required in under 2 seconds.

Requirements also can be distinguished by which stakeholder they are most important to. For instance, the example of ‘requesting a patient history’ may be more important to patients and healthcare workers, whereas this may not be critical to the government’s disease control needs.

For the possible functional requirements, *use-cases* (again, what the user *needs* the system to do) can be extracted from the literature. Examples of common use-cases requested in a healthcare network are shown in Table 2.

Table 2: Common use cases in a healthcare network

#	Use case	Primary	Secondary	Stakeholder
1	Scheduling appointments, referrals or tests		✓	Healthcare worker
2	Medical reports on patient	✓		Healthcare worker
3	Survey of medical facility		✓	Government
4	Electronic prescriptions	✓		Healthcare worker
5	Request for information: lab results, care reports	✓		Healthcare worker
6	Access medical knowledge base		✓	Healthcare worker
7	View and update patient history	✓		Healthcare worker
8	Transfer a patient in/out from/to another facility		✓	Healthcare worker / Patient
9	Patient information service		✓	Patient

The majority of these use cases are from the perspective of the healthcare worker and are generally based around data that is used for the *primary purpose* of direct patient care. However, there exist other use cases that use data for *secondary purposes* in the care process. These use cases, generally, come from the other stakeholders, for instance patients may need to have access to ‘Patient Information Services” (Beyer, 2004)or even have control over access to the data (Abiteboul, 2004), and government needs to be able to conduct national surveys collecting indicators on the performance of public health care (South African Department of Health, 2001).

In terms of non-functional requirements, the primary one for an NPR is the ability to perform the use-cases above *across all medical facilities in the country*. This is a non-functional requirement regarding the *availability* of the system. This requirement would be relevant to all stakeholders in varying degrees. Other non-functional requirements include: *performance*, the need for response times to be minimized in patient searches (Ohio State University, 2001); and *security*, where all access, as well as transfer, of data must be secure and privacy conscious(Abiteboul, 2004)(Northrop Grumman, 2006).

These follow the general theme of non-functional requirements, including a selection of desired properties of the system, that the stakeholder holds as most important. In general, non-functional requirements are chosen from the following list as most important (International Institute of Business Analysis, 2008, p. 198)for a particular system:

> Performance (quick response time or throughput), scalability, reliability, flexibility, security, data integrity, auditability, interoperability and maintainability.

2.2 National Patient Record Systems

2.2.1 International Efforts

Globally there is much investigation into electronic patient records and their use on a coordinated national scale. National systems to support healthcare are already in use in some countries, most notably Canada, Denmark and New Zealand. Figure *3*(from the Accenture survey mentioned earlier (Accenture, 2006)) shows the extent of Electronic Health Record (EHR), another name for an EPR, adoption into the national healthcare of countries globally.

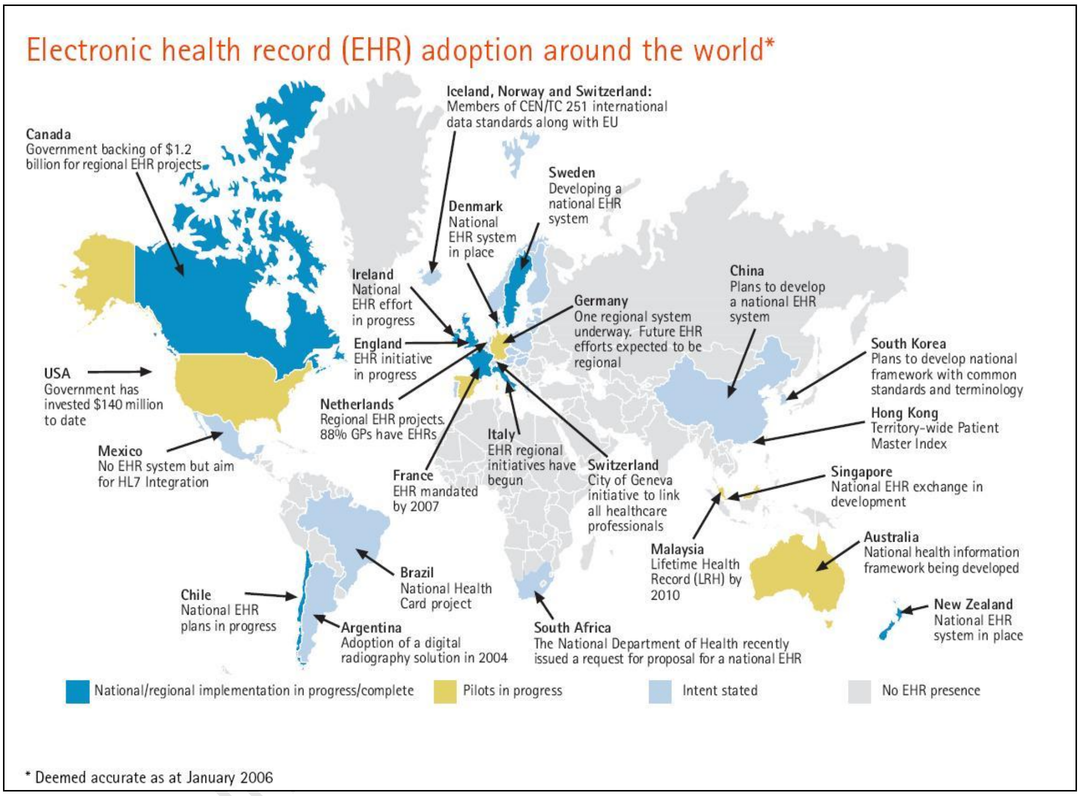


Figure 3: EHR adoption globally

Interestingly, the solutions attempted in each country, as well as the associated progress and success, differs greatly from country to country.

In New Zealand a National Health Index (NHI) system is used to support the location of patient data across the country. The system has been in place since 1992, and while in use, the system continues to be enhanced (New Zealand Health Information Service). Alternatively, Britain has attempted to support the entire National Health Service (NHS) with plans to introduce several applications that run on top of a national network for healthcare systems. Known as 'The Spine', it is planned be a nationally available EPR system that local applications can use to load the latest data on a unique patient (Ringholm, 2007).The program as a whole, known as the National Programme for IT (NPfIT), requires an enormous effort and started with a reported budget of GBP 2.3 Billion. Since then, the programme has received heavy criticism for being behind schedule, over budget and, more recently, concerns have arisen over poor security (BBC News, 2009).

In the United States (US) the government called for the development of electronic health records for most Americans by 2014. From this, the department of health and human services announced plans for a National Health Information Network (NHIN) to support nationally available electronic health records (5). Four consortia have been awarded contracts to build prototypes for the national system. Rather than unify all under a single national system (as with the British NPfIT approach), the NHIN will look to be an 'extension of local success into a nationwide network' (Northrop Grumman, 2006), meaning that it will attempt to co-ordinate with health information repositories that currently exist locally. However, the lack of nationwide standards remains a barrier to facilities sharing information. This requires much effort to formalise a nationwide standard, as well as considerable effort to extend the software at each facility to incorporate the new standards.

Despite their varying levels of development and success, the fact that these nations have started investigating solutions indicates the potential value of such systems. However, the Accenture survey into EHR adoption (Accenture, 2006)notes that: "*Among all 6 continents progress investigating the potential of electronic health records has been slowest in Africa and South America*".

To gain a better understanding of the how the various international efforts are approached, the following section will look at two of the biggest national systems that are currently being planned or developed.

2.2.2 National Health Information Network – US

As discussed earlier, the American government has commissioned a National Health Information Network (NHIN) linking healthcare facilities to a national network that will allow interoperable data exchange.

Contracts for designing and developing prototypes for the NHIN architecturewere awarded in November of 2005(U.S. Dept. of Health & Human Services, 2005). Four consortia, led by Accenture, Computer Sciences Corporation (CSC), International Business Machines (IBM) and Northrop Grumman respectively, were contracted to build a working prototype of a NHIN solution that would adhere to certain core requirements and provide proof of concept by each connecting three health care districts.

Contextual Influences

The US healthcare system is one that is highly fragmented and disparate. Most states in the union have their own healthcare legislation that builds upon the framework of the national (often referred to as 'federal') law, and this leads to differences in healthcare laws and procedures in each state. Furthermore, private medical insurance is heavily intertwined with healthcare in the US, with 85% of the nation having medical cover(Wikipedia). Many hospitals are privately owned and are thus independent of the remaining government or public healthcare hospitals.

The issue of fragmentation is not the only challenge in US healthcare, as the sheer size of the nation also makes implementing national projects (such as the NHIN) difficult. Currently, according to the US census office(U.S. Census Bureau, 2005), there are 7569 hospitals in the US serving a population of almost 300 million(CIA, 2007).

For these reasons, the four contracted consortia were given specific guidelines and constraints that would ultimately guide or affect the possible design decisions for any prototype (5). The most obvious are the set of functional requirements, which centre on the sharing of data between facilities, rather than from a centralised body. These constraints dictated an architecture that is distributed and based on interoperable message passing.

Functional Requirements

The National Committee on Vital and Health Statistics of the US government drafted the following functional requirements as the initial needs for the NHIN project(National Committee on Vital and Health Statistics, 2006):

1. Certification, authentication and authorisation of entities in NHIN
2. Personal identification – identity correlation to identify unique patients
3. The location of health records on request
4. Transport standards – use of recognised transport methods
5. Data transactions – automated between entities on trigger events
6. Audit and logging – need for accountability
7. Dynamic data access – 'time sensitive' responses
8. Communications – in standard content and message formats

Northrup Grumman, one of the bidders, consolidated these into the following functional requirements:

1. **Translation –** messages need to be converted into a standard structure and use a set of standard reference codes for terminology.
2. **Information modelling –** internal communications in the network must be translatable to the canonical data.
3. **PatientIdentification and Location –** identify unique patients from demographical information provided and manage retrieval of records.
4. **Message routing –** destination of the various communications needs to be selectively routed.
5. **Auditing –** all transactions, state changes and access to the information needs to be recorded and traceable.
6. **Confidentiality –** Control access to and secure data at rest and in motion

Non-Functional Requirements

The IBM consortium’s plan for the NHIN prototype declared that the NHIN solution must be:

1. Private
2. Secure
3. Seamless
4. Flexible, Open, Transparent
5. Responsive
6. Reliable
7. Affordable
8. Simple
9. Scalable

Network Architecture

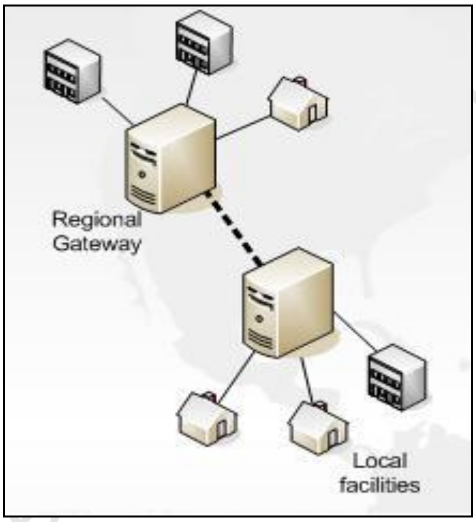


Figure 4: The decentralised peer-to-peer topology of the US National Health Information Network

The NHIN architecture uses a *decentralised peer-to-peer topology* with *super nodes*. As illustrated in Figure 4, the super nodes are regional gateways that provide a local group of facilities with access to the rest of the NHIN network. The regional Gateways are connected peer-to-peer and are responsible for resolving local facility requests for remote records. It facilitates this by translating between the national data standards of the NHIN and the proprietary data standards used at the local level. Following the functional requirements outlined by Northrop Grumman earlier, these Gateways perform four functions:

- Translation to the NHIN canonical data model,
- Location of data, and identification of patients, elsewhere in the NHIN,
- Routing of messages (sends or requests) to other nodes in the NHIN,
- Provision of security to data at rest and in transit, as well as controlling access to requests to dependent edge nodes.

Communication protocols

The edge nodes in the NHIN continue to use their current data format to handle information. When communicating with the rest of the NHIN, they send a message to their local gateway which translates the data into the national standard, known as the 'canonical data set'. The canonical data set is a set of identifiers decided in the national standard, and all gateways, convert their local formats into the canonical set to process requests or queries. Benefits of this are that the gateways assume all the burden of translation and that if the standards are changed the rules need only be changed at the gateways – not at all facilities nationally. A drawback to this, however, is that the gateway needs to be constantly up to date with whichever data formats are used in its region.

2.2.3 National Program for IT – UK

In October 2002 the UK government established the National Program for IT (NPfIT), an effort to support the National Health Service (NHS) with a number of IT applications, including a national electronic patient record.

Functional Requirements

A number of applications are planned to be included in the project, including: Choose and Book, an electronic booking service; an Electronic Prescription Service as well as HealthSpace, which allows patients to view their own health records.

These applications are to be built on top of the NHS Care Record Service, which is also known as 'the Spine'. The Spine is a nationally available network that allows access to unique patient data, including demographics, clinical data and data for secondary use.

Specifically, the Spine will be required to (Ringholm, 2007):

- Find a patient's record from demographic data (using data matching algorithms).
- Return a patient's record on request.
- Route application messages.
- Control access to records by patients, legitimate care workers and other applications.

Non-Functional Requirements

The non-functional requirements for the UK system were more difficult to collect due to the difficulties experienced during its development. Security issues have been extensively highlighted and much of the programme's current publicity is focused on security and privacy reassurance. Other non-functional requirements, such as the need for quick response times from the Spine as well as the importance of data integrity, have been publicised.

Network Architecture

The UK's Spine is implemented using a centralised network architecture that provides all facilities nationwide with access to a single repository of EPRs (illustrated in Figure 5below). The Spine's network infrastructure is dedicated for NPfIT applications only, and every remote facility must be integrated into the network (Ringholm, 2007).

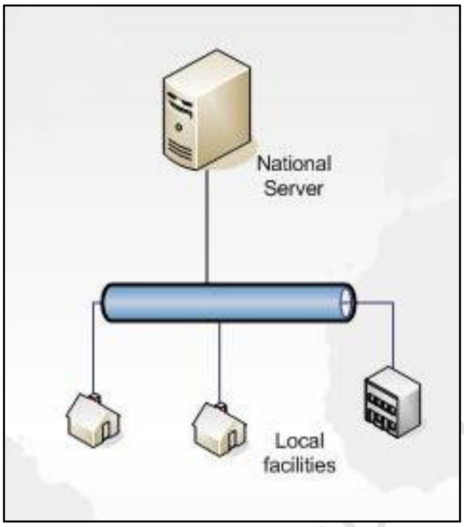


Figure 5: The centralised architechture of the UK's Spine Network

Communication Protocols

HL7 version 3 was the originally intended standard for data communication between local systems, applications and the Spine(National Health Service, 2006). However, the complexity of the data standard has caused the programme to revert to a somewhat simpler standard (that uses human readable text as well as coded data), causing yet further delays and bad publicity for the programme.

Contextual Influences

The NHS includes around 29,000 General Practitioners (GPs), 18,000 dentists and around 1,600 hospitals. The NHS itself employs more than 1.3 million people (CIA, 2007). The National Health Service also is having its own national broadband network laid, known as N3, to aid the delivery of the new IT programme.

2.3 Distributed Design of National Patient Records

Any National Patient Record system, including those outlined above, will rely on a distributed system architecture and design. This section synthesises some of the critical design aspects, as well as points of difference and controversy, that occur in the creation of such national networks.

2.3.1 Network Architectures

Any basic network consists of nodes and communication paths (Balsamo, 1998). The nodes are the points where information systems reside and communication paths are the connections between them. Nodes may be of differing types, usually based around whether they are service providers (traditionally known as *servers*) or service consumers (traditionally known as *clients*). In the analysis of the American NHIN, the gateways are servers that provide message translation, message routing and patient location services to the clients in its region.

At the highest design level, networks generally fall in to one of two categories: centralised or decentralised. *Centralised* networks consist of a single, central server that provides all services to all other members (therefore clients) of the network. Any communication between two clients has to pass through the central server. *Decentralised* networks allow for direct peer-to-peer communication between nodes, allowing each member of the network to act potentially as both a client and a server. While they may, on occasion, such as the US NHIN network, have “super nodes” or nodes that facilitate the peer’s communication, they predominantly do not have a single point that controls all communication.

An NPR using the centralised network approach (such as the UK’s Spine architecture) often contains a repository of all nationwide patient records. This approach presents many benefits, including:

- **The use of a single national communication and data standard.**As all communications must pass through the central server, all peripheral clients have to adhere to its standards for communication and data, and are unable to use others. This can also be managed ‘top-down’, meaning that additions to the standard can only be made to clients once they are in place at the centre, allowing the governing body to have close control over standards.
- **Easier analysis of the national data set.** As all data is accessible from the central server, surveys and analysis of the data is easier to perform and will only have to incorporate data from a single standard. This also can be potentially accessible in real time, creating powerful opportunities for monitoring emergency outbreaks or epidemics.
- **Once an update is made, it can be seen nationally.** If a patient’s record is updated to the central server, all other client nodes are able to view the change.
- **The Master Patient Index can be more easily maintained, and reduced risk of duplicates.**In a centralised solution that consults the server when trying to add a new patient, it is possible that a master patient index (MPI) could be maintained throughout the system. As there is only one record for a patient nationally available, and access to it is managed by the central server, contradictory duplicates are less likely to occur(New Zealand Health Information Service). Alternatively, with distributed systems, it is more likely that conflicts and duplicates would need to be resolved.
- **Decision support available nationally**. Any decision support, such as automated health monitoring, implemented at the central server will be available nationally and not have to be customised for each facility.

- **Easier to monitor and maintain**. As there is only one central server, it will be easier to monitor its performance and perform maintenance.

However, the UK's NCRS Spine follows a centralised architecture and has proved to be immensely difficult to move from a good design theory to a successful implementation. Similarly, the centralised approach can also present a number of difficulties in running an NPR, when compared to a decentralised network design:

- **The inherent danger in relying on a single server.**Problems such as power failure, viruses, software failure and maintenance updates could potentially cause downtime and halt the network as a whole. There are suggestions that many of these risks can be avoided through management of backup schemes, centralised clusters that load share, store and forward techniques and other contingency measures. However, any additional unforeseen fault in the backup systems will result in the entire national network being unavailable (King, 1983). In a decentralised network however, the failure of a single node usually only has a partial affect on the network, as only the nodes connected to it are affected. The rest of the network should still be able to communicate, albeit with one lesspeer. This also has the added benefit of the reduction of costs for backup schemes.
- **Load.**Handling real-time updates and queries from all healthcare facilities nationally is a considerable burden in any country. All systems have limited capacity in terms of amount of data and service throughput, and considerable investment would be required in the infrastructure to support this. In contrast, peer-to-peer networks rely predominantly on the local processing power of each node, and investment is only required in the raw network infrastructure to facilitate communication.
- **Local control diminished and local characteristics overlooked.**All nodes in a healthcare network will not provide the same service or use an NPR in the same way. Clinics, pharmacies and hospitals, both private and public, will have differing needs and capabilities. Research and specialised care facilities broaden the spectrum of diversity further (King, 1983).

2.3.2 Network Communications

Network communications are the second critical aspect to consider in the design of a national system. Communication protocols and standards are relied upon to keep the services of the network accessible, accurate and timely.

The architecture chosen for the network obviously impacts and ultimately defines the type of communication in the system. A design decision that comes up in most NPR systems is: how often does the local or national copy of a patient record get updated? Communication between nodes in any network can occur in one of three ways (Beyer, 2004):

- **Delivery on request.**
  Local copies of the record are only updated from the national server when a user wishes to look at the record.

- **Publish - subscribe.**
  Local copies of the record are updated from the national server whenever the record is changed elsewhere.
- **Periodic synchronisation.**
  Local copies of the record are updated from the national server at regular intervals and not when updates happen. This approach reduces the number of transactions over time and is more robust to external failure; however there is a trade-off between update costs and freshness of data.

Planning for communication failures is a critical factor in the design of any distributed system. The communication schemes above are all dependant on knowing that the messaging between nodes has occurred successfully, in order for the data to remain consistent across the network. A classic problem is that of partial failure (Arnold, 2002), where a message is sent between two nodes and a failure occurs. The sending node is unsure whether its original message has failed to arrived, or if it did arrive but the acknowledgment from the receiver has failed. Ultimately, the problem can only be resolved once the two nodes are able to communicate (or possibly by the intervention of an overseeing third party). In a healthcare context, knowledge of complete processing of a request can be crucial, such as the case of ordering drugs or procedures for a patient.

One attempt at ensuring the integrity of communication is the inclusion of *transactions* into the distributed system design. A transaction is a message that requests the manipulation of data at the receiving node, and has controls that, in the case of failure, will ask all parties to *roll back* their data to the state before the message was sent. Thus the partial failure problem is addressed – as transactions either go through completely or not at all (known as the *atomic* property of transactions)(Microsoft).

Along with failures, connections also experience certain constraints on the speed at which they can transmit messages. Latency is the delay that occurs in a message passing between two nodes in a network. While in local area networks this time is often negligible, in wide area networks this time can be considerable enough to have an effect on the fundamental performance of the system. The design of nationwide systems such as NPRs have to take latency into account when designing the communication schemes of the network, as failure to do so can result in unusable performance from the system.

In summary, each network design needs a set of rules and communication standards that allow the nodes to communicate within these various constraints. Known as *network protocols,* they are at the core of the design of any distributed system, as all nodes must adhere to them.

2.3.3 Distributed System Data

Once the architecture and communication protocols of a network have been designed, the nodes are now able to communicate information, which is usually the primary goal of the network in the first place. As most systems will rely on some machine-interpretation of the data passed, the expectations of what the data represents need to be agreed upon.

Messages passed in any NPR have to obey a common standard that define two things: the syntactical logic of the message (the 'encoding'), and a standard of common terms or codes in which the information is described (the 'vocabulary'). The task for translating user or system-specific data into an interoperable format can be done at different levels – at the server, at the client or at an intermediary. The server can accept messages in all formats and on receipt translate them into the standard data; however this is the most unfeasible option due to the burden on the server. More commonly, clients are altered to translate their data into the national standard (as planned in the UK's NPfIT(Ringholm, 2007)). This requires all parties who wish to use the national server to ensure they translate their data correctly. An alternative to both of these strategies is the use of an intermediary node in the network that is able to understand both client data and the standard for the national network. This approach can be seen in the US's NHIN prototypes(Gartner, 2007), where regional 'gateways' translate local data into the NHIN's canonical data set.

Another issue regarding translating data is: at what level should data be standardised? Data standards do not necessarily have to apply to all data that is shared, as it would be possible to pass plain text or documents between nodes without having their contents encoded. The argument for codifying all data is that it allows all data to be understood and analysed by a computer and is thus easier to provide statistics and use automated rules. However, if a facility is not able to change their current system to adhere to the codified data standard, they are unable to participate in the network. This has resulted in Bayer et al. (Beyer, 2004) suggesting multi-layered communication and data standards, where servers will accept data at different levels. In this work they suggest the design of a central repository that can accept data at three different levels:

1) Data that is completely codified to some specific ontology or standard.
2) Data that has metadata tags to categorise or summarise the message which contains human readable text.
3) Data that has no codified data (only patient and provider details), and is simply a file that can be stored in the repository for access by other parties.

One of the predominant design issues for NPR systems is the data standards around patient identification. Naturally, in an NPR, probably the most common functional requirement is to find a unique patient's record, which is usually found using either the patient's name or a pre-assigned unique identifier. Coordinating the creation of unique identifiers for each patient in the population was often one of the most difficult parts of implementing such systems. Recently however, many systems have moved towards technologies that use algorithmic matching of patients on demographic data, to find the correct patient record. This method is certainly easier to co-ordinate, although questions still remain on the impact of the time it takes to make a match as well as the impact of false matches (despite the apparently low probability of this happening). It is interesting to note that both the NHIN and NPfIT plan to use the statistical matching method(Ringholm, 2007)(Gartner, 2007).

2.4 The South African Context

The South African government has indicated that they plan to implement a *National Health Information System* for sharing and collecting patient records on a national scale (South African Department of Health, 2001). One of the key features of this system would be a National Patient Record (NPR) that would be accessible across the country. The following section will outline the unique contextual challenges that will impact any model for a South African NPR. While requirements and technical approaches can have commonalities between countries, it will be the system's applicability to the local context that will determine the success of any NPR.

South Africa has a complex context of contrasting extremes for implementing an NPR. On the one hand, there are areas of relatively good ICT infrastructure, while on the other it has major challenges with general healthcare infrastructure and capacity. South Africa's teledensity (total telephone subscribers per 1000 people) is comparable to most middle income countries and is four times the average for sub-Saharan Africa (SSA) countries (World Bank, 2006). Though South Africa's number of trained health workers is better than most other countries in SSA, it suffers from the fact that while the majority of the population uses public healthcare, the majority of healthcare workers work in the private sector(Health Systems Trust, 2005), leaving public facilities understaffed with a high patient to healthcare professional ratio.

There are many differences in the operating environment of a developing country such as South Africa, and the developed countries from which most of the NPR designs have emanated thus far. The following sections will consider South Africa's context in terms of: scale, node context, load and other barriers.

2.4.1 Scale

The majority of healthcare facilities in South Africa fall under three categories: public hospitals, private hospitals and (public) primary health care clinics (PHC). 396 public hospitals, 357 private hospitals and around 3182 PHCs serve South Africa's population of 48 million people (Health Systems Trust, 2004). Facilities are distributed differently throughout the nine provinces and, despite recent improvements; poorer provinces tend to have fewer facilities and healthcare professionals.

2.4.2 Node Context

There are many disparities between healthcare facilities in South Africa. For instance, there are 357 private hospitals and 396 public hospitals - however, 82% of the population relies on public health care. The fact that 72.6% of doctors and 41.1% of nurses work in private health care shows that the disparity in resources can be considerable (Health Systems Trust, 2005). Urban versus rural facilities also show disparities in resources, as shown in the 1998 health review (Health Systems Trust, 1998):

> ‘*With regard to the distribution of health sector personnel, disparities tend to occur between the more urbanised and historically better funded provinces and those which are predominantly rural. For example there are forty nurses per 10,000 population in the Western Cape whereas in the Northern Cape and Mpumalanga there are only 20*’.

Thus, in turn, the province a facility is in also can affect its resources – depending on whether it is a generally urban or rural province.

2.4.3 Load Characteristics

The load at any node in the network will be dependent on the amount of activity at the facility it represents. Activity at a facility is usually represented by the population that is treated at the facility and the frequency with which they visit. The South African Health Review, provided by the Health Systems Trust, provides vital data to calculate these two numbers (Health Systems Trust, 2004). Firstly, the population of each district is reported along with the average number of visits to a healthcare facility per person per year. While the national average is 2.3 visits to a healthcare facility per person per year, this number can fluctuate across health districts from less than one visit (Metsweding District, Gauteng) to more than four (Frances Baard District, North West) on average for the district. Again, the load for the different facilities will depend greatly on whether the facility is public or private, urban or rural, as well as which province it is in.

2.4.4 Operational Barriers

Barriers to the successful implementation and usage of an NPR can include issues around system usage, national infrastructure, regional politics as well as gaps between the theoretical design of a system and its implementation in reality.

The usage of a system by its end users, in this case healthcare workers, can make or break any system. Once users stop using the system, or stop using it correctly, the system becomes redundant and is often difficult to resurrect. In South Africa, both computer and written language literacy can be a challenge in the use of systems by unfamiliar users(Heugh, 2007). Furthermore, this illustrates the importance of ‘buy-in’ to a system – the sense of ownership users often build with a system they perceive to be valuable. Buy-in can be severely affected by quality issues such as performance speed, data correctness as well as usability.

Accuracy in captured data also can greatly affect the success of a system – as its results will only be as good as the data captured. Accuracy in data ties in closely with issues like usability and system buy-in. However, users need a sound sense of the reasons for the data capturing; otherwise apathy can develop towards the data capturing process.

As discussed earlier, resources at a healthcare facility often depend on its private or public status and which province it is in. In low-resource settings, hardware availability, reliable power supplies and access to phones or network connections often are constraints. Even when these are available there can be tremendous difficulty in finding support for both the hardware as well as the users of a system.

Political wrangling can destroy the best of efforts and their potential influence over a national system is obviously great. However it is stated by (Braa & Hedberg, 2000)that South Africa has a good environment for change and new developments. Despite suffering from ailments such as political power struggles and corruption, Braa et al. insist South Africa is not unique in these troubles and is in fact far better positioned for progress than similar developing countries such as Cuba and Mozambique.

2.5 Research Approaches for Evaluating Networks

There are several approaches available for evaluating network designs. Often the techniques used are dictated by the specific network being planned. If it is a fairly common network problem, network designs can be produced from previous practical experience. However, when considering novel network designs or, as in this case, considering an existing network design in a *new* application context, some research design is required to answer key questions about how the new network is likely to behave. Traditionally this research includes three possible approaches: the use of analytical models to calculate a section of network behaviour using probable values, the use of experimental system development to gain insight into the real-world challenges and, lastly, the building of a simulation model in an effort to replicate the real world behaviour. The next sections will consider each of these approaches and their applicability to the presented problem.

2.5.1 Analytical Methods

Analytical methods of evaluating networks are characterised by the use of mathematical models to represent systems and that uses analytical calculations on the model to predict system behaviour. Most models will require parameters to describe the initial conditions, and can often make use of classical mathematics to incorporate the complexities of random events in the networks.

Analytical methods can be cheaper to implement and require less investment than the alternatives of experimental systems or simulation. Furthermore, analytical methods are generally faster than alternatives, as they do not require as much effort to build supporting tools and frameworks (Molkdar, Burley, & Wallington, 2002).

Various mathematical techniques have been applied to network evaluation. The Markov process and its stochastic capabilities have many applications in calculating the probability of events occurring. Other mathematical tools such as Bayesian networks, Fault Tree Analysis (Franke, 2009), Monte Carlo calculations (Molkdar, Burley, & Wallington, 2002) and the Queuing Network Model (Balsamo, 1998)have been used in network evaluation. The models can address various themes, such as the efficiency of message passing in the network, bandwidth demand and expected error rates.

However, it is seldom the case that one study, or technique, can produce results or information on more than one problem at a time. Predominantly, key process or events need to be modelled and analysed in isolation due to the restrictions of the mathematical models. As such, there is a trade-off between the rapid speed at which analytical methods can be used, and the coverage of the model or design being evaluated.

Furthermore, analytical models tend to be poor tools for communicating or demonstrating the behaviour of a network or its intricacies. They tend also to produce less accurate results as they require more simplified models to remain feasible. These factors normally make them a poor candidate for planning new network designs, and then rather tend to be used for evaluating specific problems in established networks(Molkdar, Burley, & Wallington, 2002). Indeed, it can be proposed that they are used as a first pass to do scoping, then other techniques can be used to provide a more thorough analysis.

2.5.2 Experimental Methods

The methodology for experimental network evaluation consists of proposing a network solution based on contemporary knowledge, and then implementing scaled-down versions of the network in the operating context. The implanted system is then carefully monitored on all performance metrics and can have its data reviewed periodically (50). This approach is often favoured as having a mini-model in the actual operating context, is the most likely method of finding out what the practical and technical barriers to the solutions larger operation are.

While this certainly is the best method for gathering data on real world performance, such experimental methods are characterised by being extremely expensive to implement. For NPR networks, in particular, this approach is unfeasible on a national scale. This explains somewhat the common practice of running “pilot” implementations with smaller subgroups of the target network, to evaluate the designs performance in the real world. This was the approach of the US NHIN project that tendered the NHIN architecture to 3 consortiums to each develop a prototype of the system in a specific region.

However, the expenses of this still can be immense. Mistakes in architecture design or incorrect assumptions about the operating environment can have catastrophic impacts, and require considerable resources to remedy. The best example of this is the UK NPfIT, which has suffered many unforeseen challenges since its attempted national rollout (17).

2.5.3 Simulation Studies

Simulation studies are somewhat of a middle ground of the two preceding approaches. A conceptual model of the network to be evaluated is created in a computer program that aims to replicate its real world behaviour. Much like analytical methods, parts of the simulation depend on logical algorithms and rules to govern certain component’s behaviour but, also like the experimental approach, simulation is able to combine many complex parts and attempt to provide a realistic environment for them to interact naturally.

To date, simulation has been used in healthcare to mostly model either epidemiological scenarios, such as the spread of disease in a population, or to analyse process efficiencies, such as lab test ordering or processing queues in health facilities (Au-Yeung, 2006).

Simulation has been used widely in the network research field to evaluate various problems and scenarios of new network protocols or technologies, but only a few efforts have focused on using it at the abstracted national network level. Work by Dudenhoffer (Dudenhoffer, 2002) proposed the need to use simulation to model the impact of different network emergency crews. Dudenhoffer goes on to suggest that as simulation was able to include many more real world components and influences in its calculations, it was potentially able to observe the macro emergent effects of the modules interacting. Such insights could not be from the simple analytical models, and recreating such complex worst case scenarios in an experimental effort was infeasible.

Simulation also is a compromise of the alternative approaches in terms of cost and time. While as not as rapid as analytical methods, there are many simulation tools available that allow a simulation to be built relatively quickly when compared to the experimental approach of building the system. For similar reasons, it can also be seen that the costs of simulation are less than the experimental approach, but would normally be more than the analytical approach.

However, large scale simulations such as that of a national network (like an NPR) also face a set of challenges. The model has to incorporate an increasing number of factors as the size and complexity of the network grows. Furthermore, the amount of events generated requires adequate computational resources for the simulation as well as the analysis of the data (Waupotitsch, 2006). The simulation of such large networks (such as global Internet usage) remains a Grand Challenge in network simulation. Waupotisch and Eidenbenz (Waupotitsch, 2006) propose that new approaches to simulation are required to provide solutions that can: model solutions end-to-end, which increases the realism of the model; are flexible enough to allow various kinds of network component; and provide new approaches to scalability.

Ultimately, the quality of the outputs of a simulation study has a strong correlation with the investment in the research (Sargent, Model verification, 2007).

2.5.4 Network Simulation Tools

Today simulations do not have to be built from scratch, as there are many tools to support this. The main function of simulation tools is to allow researchers to focus on creating the model that is being investigated, and not need create the scheduling and framework to run the simulation.

Any simulation needs certain elements: a model, input parameters and a simulation run. The model is the representation of the real world thing, such as a network or a system, that is being evaluated; the parameters are input values that customise the behaviour of either the model or the simulation, for example node failures. Simulation runs are when the a particular model is tested against a set of parameters for a specific duration. More specifically, parameters are used to specify three main things in the simulation; the duration of the simulation (an hour, a year), the schedule of events occurring in the simulation and also any constraints or capacity of the simulation. Timing is also important in a simulation, as simulation tools can support these models either by discrete (event based) or continuous (real-time) representations of time.

Two common tools used for network simulation research include the open source OMNet++ (Varga, 2001) as well as NS2 (Fall & Varadhan, 2004); as well as several other commercial options including OPNET and others.

2.6 Summary

This chapter has considered the various themes around the design of a National Patient Record solution in South Africa, as well as techniques for its evaluation.

Globally, the benefits of Information Management in healthcare are being recognised and more countries are considering the implementation of a national system for sharing Electronic Patient Records. Two predominant international efforts are those of the US and UK. The US's decentralised network architecture and the UK's centralised approach present different implications for the design of communications, data, services and security of the solution.

If South Africa considers either of these designs for its own NPR system, the unique differences in its operational environment to that of the US and the UK could affect the efficacy of the NPR. Each of the solutions would have to be evaluated to see how it responded to the challenges of rural-urban disparities in access to healthcare, connectivity issues and other human barriers.

Simulation proves to be a strong candidate method for performing this evaluation over the less powerful analytical methods and the more expensive experimental methods. The following chapters will demonstrate how this background information on the NPR solutions and contextual environment contribute to the design of a NPR simulation model that is used to evaluate the performance of the US and UK networks in the South African context.

Chapter

3 Modelling and Implementation

Given the established problems that face implementing a national EPR system, simulation was proposed as a method to evaluate the various technical network approaches in specific environments.

This chapter will explain the methodology followed to evaluate the network architectures, and how the simulation tool OMNeT++ was used to do this. Using the available literature, a model of healthcare systems and their operational environment is presented, as well as its programmed implementation in the simulation tool.

3.1 Research and Evaluation using Simulation

This work follows the standard evaluation of systems using simulation, as defined by LAW (Law, Designing a simulation study, 2003), with the major stages being:

1. Formulate problem
2. Collect data and define a model
3. Validate model
4. Construct simulation
5. Validate simulation with pilot runs
6. Design experiments
7. Make production runs
8. Analyse output data
9. Document and implement results.

Of the stages listed above, this chapter will formulate the problem and goals for the simulation, demonstrate how the collected data was used to define the model and explain the construction of our simulation. For the other stages, chapter five will address the validation of model and simulation, while chapter six contains a case study that demonstrates the experiment design, production runs and analysis of output data.

OMNet++ was used as the simulation tool for this research. Like other available tools, it allows the definition of a network model and the simulation of its behaviour over a discrete set of events. This allows the research to perform the key experimentation steps from the list above, from constructing a model through to producing experimental data.

The rest of the chapter will explain the building of a model within the OMNet++ framework.

3.2 Formulating the Problem

As introduced by the research question stated earlier, the core problem being investigated is:

> *"Is simulation a feasible technique to evaluate designs for a National Patient Record in South Africa?"*

To address this problem, the research began by investigating the different network architectures that are currently in use for an NPR. Further work analysed the various metrics made available by network simulation and examined the challenging context of developing countries.

3.3 Defining the Model

The NPR network model has to represent various modules of the real world. Some modules are technical in nature, such as network connections or file sharing algorithms, while other modules have a contextual representation, such as a population's usage of the healthcare system. For each of these modules, the following sections will describe their motivation for each module's inclusion using real world examples and how they are modelled and simulated using OMNeT++.

3.3.1 Description of Modules

Any network consists of three core elements: nodes, connections and messages (Balsamo, 1998). A Network Architecture is the connection of nodes in a specific arrangement that allows messages to be passed directly between particular nodes.

In the healthcare networks studied, nodes will typically be either a healthcare facility or a 'server' node that support the technical architecture allowing facilities to communicate. An example of this is the UK NpfIT system, where any facility (a clinic, a hospital or dentist) can request a patient's history from the national unified server (Ringholm, 2007). The NPR model developed recognises this by having to basic kinds of node to represent the technical network: either a facility or a server. The population also serves as a node as they have a clear relationship to facility nodes, and a patient visiting a facility can be seen as a message being passed between the population node and the facility node.

Connections predominantly represent physical network connections between two systems that are geographically separate. As an example, a PC in a clinic accesses the national repository over a dedicated connection. Figure 6 below illustrates the connections between server and facility nodes, and also demonstrates the role of the population as a node with connections to each facility.

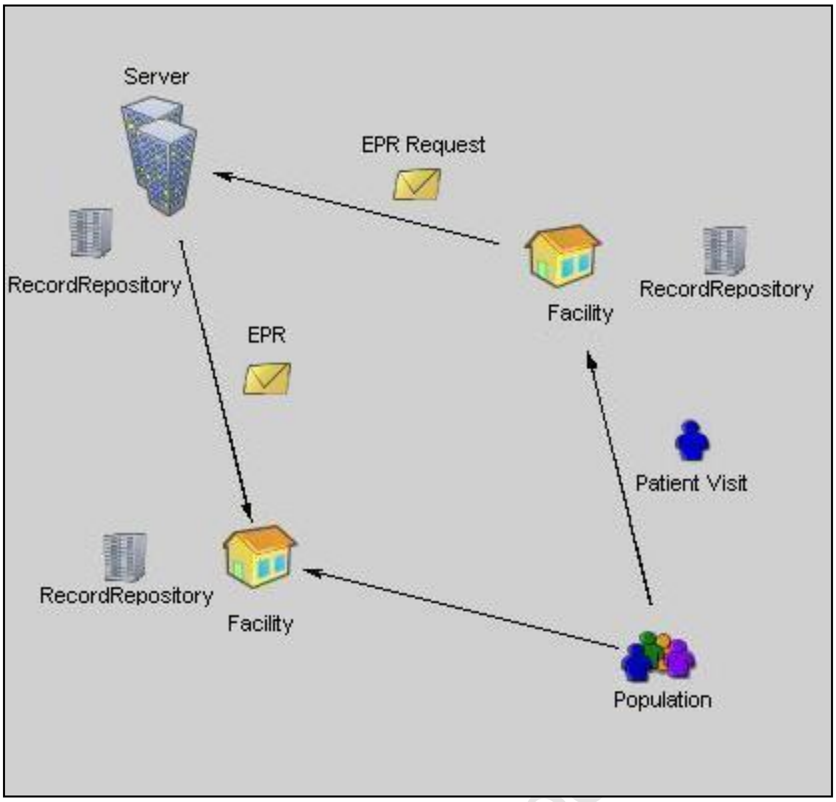


Figure 6: Basic Network Modules

Messages represent information passed between two nodes in a network over a shared connection. Often, in network simulations, messages are used to represent the individual packets of data that make up a greater message being passed in a particular network protocol. Alternatively, they can represent the transaction as a whole. In this case, the example of a PC accessing the national repository would involve two messages: one as the request message travelling from the clinic's PC to the national system, and the second as the EPR sent as a response. This is illustrated in Figure 6 by the directional arrows, as well as the use of messages to represent individual patients visiting clinics from the general population.

A RecordRepository also is added to the model to represent the various data structures needed to maintain the EPRs at each node, and is discussed in detail below.

The following sections will describe each of these modules in further detail, and elaborate on their specific programmed implementation in the OMNeT++ tool. To provide consistency, the modules that make up the model will be described using the following structure:

- ***Module description***- the role of the module in the network.
- ***Motivations from the real world*** - the real world scenarios that led to the inclusion of the module in the model.
- ***Implementation in OMNeT++*** **-** the design of the module in OMNeT++, specifically describing the module's structures, connections, messages and any parameters that can be used to tailor the behaviour of themodule.

3.3.2 Population

Module Description

All EPR transactions in the network ultimately stem from its interactions with patients from the covered population. A population has a particular size consisting of individual patients. The patient (as discussed later) has individual behaviours and initiates individual patient visits to facilities. Thus, the population has direct access to facilities in the network, and is the source of their primary external event. A population is modelled as:

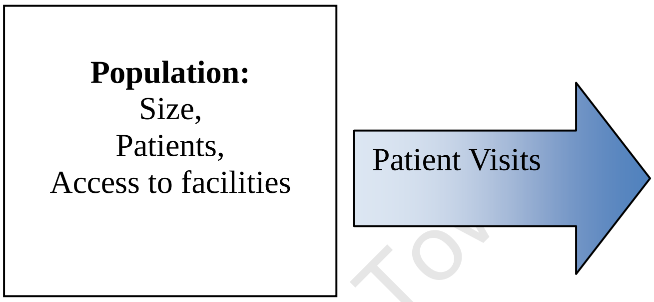


Figure 7: Population Model

Motivations from the Real World

Examples from the real world include South Africa's Western Cape Province, which has a population of around 5 million people, and several hundred healthcare facilities. In theory, all patients have access to all facilities in the province, although in reality this is seldom the case as factors of proximity, income, illness and others often determine the facility visited. The population's usage of the accessible facilities can be abstracted to a series of patient visit events over a period of time.

The model's characteristics are customisable and therefore also allow for a number of populations to be used in the same network model. Use cases for this could be the analysis of how different districts use the same set of healthcare facilities, or even two populations separated by some attribute such as location, gender, illness or some socio-economic variable.

Implementation in OMNeT++

The Population Module has the following implementation in OMNeT++ :

- Connections
    - Is connected to a fixed number of facilities
- Internal structure / elements
    - Contains individual patients (data structures) that each have characteristics that will determine how often they need to visit the various facilities.
    - Contains its own RecordRepository data structure of patient visits –

to serve as a master copy of patient visits, whichthen can be used for comparison with those used at other nodes in the network. Its role in this case is as an analysis tool for the simulation study, and not representing the real world object.

- External events handled
    - none
- External events generated
    - Before the simulation begins, *patient_visit* messages are generated and scheduled for each patient over the simulated period.
- Input parameters

Table 3: Population Parameters

Parameter	**Description**	**Example Value**
simDuration	The possible period for generating requests (ultimately the max duration of the simulation)	28800 s (8hours)
PopulationSize	The number of unique patients in the population	1,000,000
PatientType Distribution	Distribution of patient types (explained in the next section)	25% for each type

3.3.3 Patient

Module Description

The key source of events (protagonist) in an EPR system is the patient. Each patient is unique, and the demand and interactions with the healthcare system can differ greatly across individuals. Issues such as age, illness, location and economics can affect how often a patient visits a facility as well as which facilities they visit. However, in terms of an EPR network, an individual's various attributes ultimately result in two output effects on the system: the frequency of visits over a period and how those visits are distributed across the available facilities. Thus, a patient can be modelled as:

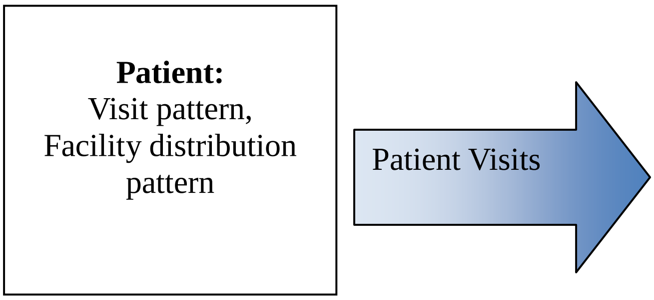


Figure 8: Patient Model

Furthermore, these two attributes of a patient actually describe the four major types of patient in the average population: acute, chronic, random and absent. The relationshipsamong these types and the attributes are shown in Table 4 below.

Table 4: Patient Types

Patient Type	Visit pattern	Facility distribution pattern
Absent	No visits in period	Not applicable
Random	A few random visits in period	Uniform
Acute	A cluster of visits in period	Majority at one facility
Chronic	Visits at regular intervals in period	Major portion at one facility but with visits elsewhere

Motivations from the Real World

This patient model can cater for different kinds of patient, aiming to cover the spectrum of possibilities that exist in a real population. To illustrate, a patient who only has a sore throat once in a year will probably result in one or two visits at almost any facility is given the type *random*. The *acute* type represents a patient who suffers from an illness that requires sudden treatment over a finite number of visits – a good example of this would be a broken leg – usually at the same facility. Alternatively, a patient newly diagnosed with a chronic respiratory disease will more likely have frequent periodic visits over the year and have visited or had tests done at more than one facility, and is thus distinguished as a *chronic* type patient.

It is certainly possible that this model could be elaborated on to include other attributes that affect a patient’s visit patterns and facility distributions, such as urban, rural or other factors. Furthermore, the relationship between the patient's distance from a facility and probability of visit could potentially be valuable additions to the model.

Implementation in OMNeT++

The patient model is implemented as a class in OMNeT++ as it is external to the traditional network elements. Its core class definition is included in Appendix A.

3.3.4 Facility

Module description

Facilities are simply a source of requests to the national system. These requests would be to either create, request or update a particular patient record. The requests result from the ability to consult (care or treat) patients, as a function of the frequency of patient arrival. Thus the facility can be simplified to the following model:

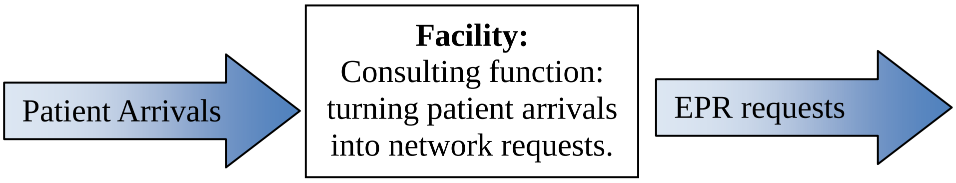


Figure 9: Facility Model

This model is able to cater for different kinds of facility. A primary healthcare clinic, a blood testing laboratory and the emergency ward at a provincial hospital will all have different patient arrival trends as well as processes for consulting and caring for patients. These two factors will determine the interaction with the EPR network.

Motivations from the Real World

This model can be illustrated by using it to represent a typical Primary Healthcare Clinic (PHC) in South Africa. PHCs normally experience visits in the order of 40 to 100 patients per day from the surrounding community. This arrival trend also can be observed to be dependent on the time of day – with the clinic often experiencing its busiest times in the early morning. Also, patient visits can only occur during the hours the clinic is open, imposing a window period on the time that the PHC will interact with the greater EPR network. These three factors contribute to the arrival model.

For the consulting function as referred to above, three processes were considered: first, how a facility deals with its arrivals; second, the process the patient goes through to receive care; and third, how the consultation results in an EPR network interaction. In the PHC example, the norm is that, on arrival, patients join a waiting queue to be seen by a doctor, with different facilities having different numbers of consulting doctors available. On getting to the front of the queue the patient can then 'be seen' by the doctor available. While the 'consultation' can last for varying amounts of time from patient to patient, the DoH stipulates that 6 minutes should be the target maximum length. So it is assumed that most visits on average take this amount of time. Critically, it is also in this consultation period that usually any interactions with the EPR network would occur, including requests for medical histories, progress tracking and prescription ordering.

It should be noted that this model also caters for 'patient proxies'. A patient proxy is an object or an event representing the patient that can cause the patient's identity or medical record to be used / interacted with. An example of this is at a laboratory, where a patient's blood sample – the patient proxy – is tested (consulted) and the results updated on the patient's EPR. The lab also adheres to the model presented above, receiving samples following some arrival trend (often in batches from a group of local clinics), and then processing the samples in some order and ultimately updating the EPR network to provide the results.

A facility's major challenge with respect to a nationally available EPR is the reliability of its connection to the network. Many PHC facilities in South Africa are currently not connected to the Internet and those that are can often experience downtime due to infrastructure instability. This nature of instability at the node was included in the Facility module – an inherent internal ability of the node to suddenly become disconnected from the network for a period of time.

Implementation in OMNeT++

The Facility Module has the following implementation in OMNeT++:

- Connections
    - It is connected to one or more population modules, from which it can receive *patient_visit* messages.
    - It is connected o zero or more Server modules.
    - It is possibly connected to zero or more other facility modules.

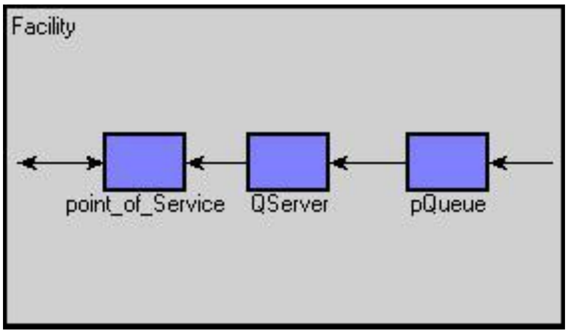


Figure 10: Facility Submodules in OMNeT++

- Internal structure
    - The facility maintains one or more *queues* of patients awaiting consultation (as shown in Figure 3), implemented using a pQueue submodule.
    - It has one or more *Point_of_service* submodules that "consult" patients when received from the queue, representing a doctor, nurse or other such healthcare worker who interacts with the EPR system on the patient's behalf.
    - For each queue, a *Queue Server*sub module is required to manage the distribution of patients in a queue to any Point_of_service modules available.
- External messages handled
    - When a *patient_visit* message arrives from the population module, the facility adds the patient to its available queue. Currently there is no algorithm for queue addition and it works using the traditional FIFO rules. The facility module could certainly be extended to include logic for perhaps assigning patients across multiple queues or even inserting a (say, very sick) patient further into a queue.
    - *patient_visit* messages are generated by another module, the *Population module*, which is discussed later. Thus the only attribute of the facility that affects the arrival trend is the hours that the facility is open in a day
- External messages generated
    - Depending on the EPR solution algorithm, the facility's *point_of_service* nodes can generate EPR network messages. These may include requests to create, request, update or find a particular EPR from the network.
- Internal events or messages generated & handled
    - If an EPR request goes unanswered, the point_of_service tries to resend the message. It is reminded to do this by internal timeout events being scheduled when the original message is sent.
    - Downtime periods also are created using internal scheduled events

(or 'self-messages') to either 'start downtime' or 'stop downtime'.

- Input parameters
  - The following parameters were used:

Table 5: Facility Parameters

Parameter	Description	Example value
Timeout	Maximum allowable delay for server response.	20 seconds
maxRetries	Maximum allowable retries following a server timeout.	3
downtimeProbability	Over a particular period, the probability of a downtime event.	0.01
Meandowntime	The mean duration of a period of downtime.	3600 seconds (1 hour)
Avg serviceTime	The mean time a patient spends in consultation with a healthcare worker in a particular facility.	360 seconds (6 minutes)
period	The period used to calculate bandwidth usage.	600 seconds (10 minutes)
Loss_probability	The probability of losing any particular incoming message to the facility.	0.05
Visit window	Amount of time in a day that the facility allows visits.	28800 seconds (8 hours)

3.3.5 Server

Module Description

In an NPR network, the patient record requests generated by a facility are sent to a server. A server is responsible for implementing the particular algorithm for sharing and routing EPRs in the network. Depending on the particular network algorithm, most servers are able to: receive EPR requests, find an EPR and return an EPR to a facility. A Server can be simply modelled as:

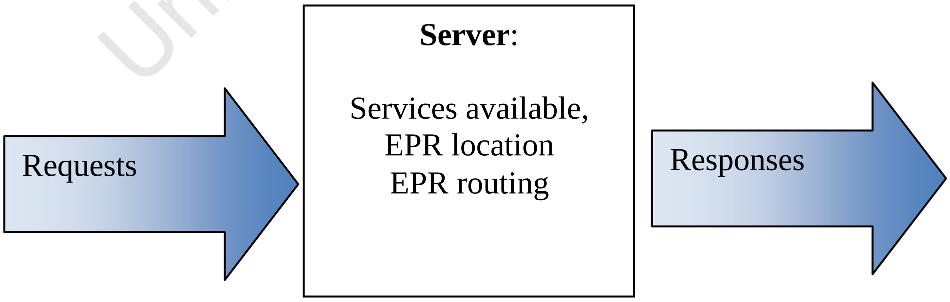


Figure 11: Server Model

This component is able to cater for different kinds of server. A server in a centralised network will operate differently to one in a peer-to-peer network, but both will be able to receive and respond to requests. Services available represent the actions that facilities (or even other servers) are able to request, such as "create EPR", "update EPR", "find EPR" and "read EPR".

Motivations from the Real World

To illustrate the use of the model, it can be considered in terms of how the servers operate in the centralised EPR network in the UK, and the plans for the peer-to-peer EPR network in the US.

In the UK's centralised architecture, there is a single server (in the terms described above), that is accessible from all facilities nationwide. Thus the server must be able to process simultaneous requests from all over the country in large volumes. The server must be able to allow facilities to access the services of "create EPR", "update EPR", "find EPR" and "read EPR".

Alternatively, in the US's distributed peer-to-peer network architecture, as facilities keep copies of the EPR, the regional servers have to focus on services that allow sharing of EPRs between facilities – both inside and across regions. Thus the services on offer would be "publish EPR update", "find list of available EPRs for patient X" and "retrieve EPR from remote facility for patient X". Crucially, this differs from the server in the UK model, as these servers also have to be able to accept requests for services from other peer servers.

Implementation in OMNeT++

The Server Module has the following implementation in OMNeT++:

- Connections
    - Is usually connected to a fixed number of facilities
    - Possibly connected to other servers
- Internal structure / elements
    - The server can have a store of patient records. In OMNeT++ this was implemented as an internal data structure (RecordRepository – discussed later) that has a hash map, mapping patient identifiers to their records.
    - The server also can maintain a directory of facilities a patient has visited or that have a record on the patient (particularly for the US peer-to-peer network). This directory can be queried to retrieve a list of facilities that a patient is known to have visited.
- External events handled (depending on algorithm).
    - When an *EPR_Request* message arrives from a facility, the server can retrieve (if available) the requested EPR from its local store and respond. Alternatively, the server can check its directory to find which facilities have an EPR for the patient.
    - When an *EPR_update* message arrives from a facility, the server can update the status of the EPR (if available) in its local store.
    - For a Server in the US model, when a *Publish_update* message arrives from a facility, the server will update its directory to include the visit of the patient to the remote facility.
    - For a Server in the US model, when a *Route_EPR_request* message arrives from a facility or a peer, the server can route the message to

the target facility.

- External events generated
    - An *EPR* message can be sent to facilities from the server, in response to an *EPR_Request.*
    - An *EPRs_found* message can be sent to facilities from the server in response to an *EPR_Request* message. *EPRs_found* would contain a list of facilities that have an EPR for the requested patient.
- Internal events generated & handled
    - none
- Input parameters
        - The following parameters were used in the Server module:

Table 6: Server Parameters

Parameter	Description	Example value
loss_probability	The probability of losing any particular incoming message to the facility.	0.01
period	The period used to calculate bandwidth usage.	600

3.3.6 Record Repository

Module Description

The RecordRepository is an internal data structure used for storing and maintaining a list of patient records. The RecordRepository keeps a unique record for each patient, which only records visits to facilities, and the version of the patient record.

Motivations from the Real World

The examples of the RecordRepository from the real world are numerous as similar data structures form the core part of any (distributed or not) EPR system.

Implementation in OMNeT++

In OMNeT++ the RecordRepository is implemented as a standalone class, independent of the OMNeT++ framework. It contains a simple hashmap for mapping patientIDs to the number of visits, and accessor functions that allow the hashmap to be updated and read from. Importantly, the RecordRepository class is not static, and can be used as a local copy in different modules in the network. A RecordRepository then can be included as a global member of different modules, be it a facility, server or population, which then allows for data freshness comparisons between the RecordRepository instances.

The RecordRepository core class definition is included in Appendix A.

3.3.7 Network Connections

Description

The role of the network connection is to pass messages between the modules listed above. The main effects of infrastructure constraints are the unreliability of nodes or connections in handling messages. If a connection is broken or a node disconnected, messages will not flow through the network as intended. Thus we the instability of network nodes is modelled as *loss probability* and *propagation delay* of any message that is sent.

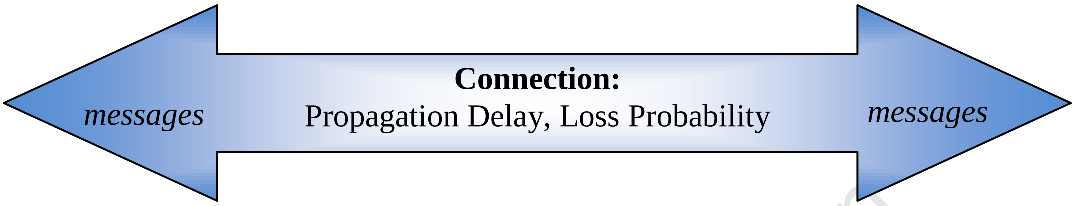


Figure 12: Network Connection Model

Implementation in OMNeT++

1. Input parameters

Table 7: Connection Parameters

Parameter	Description	Example value
propagation delay	The delay of passing a message over a connection between two nodes.	1ms
loss_probability	The probability of losing any particular incoming message to a node.	0.01

3.4 Network Model Examples

To demonstrate the flexibility of the proposed NPR model, this research included two different NPR network architectures and their associated algorithms for sharing patient records. The following section will present each of the models and document the following:

1. the network topology
2. how the main use cases are supported algorithmically
3. how the algorithm is implemented in OMNeT++
4. what additional parameters are required for the particular algorithm

Interaction with the network is stimulated by two particular events:

- The arrival of a patient at a Point_of_service – usually resulting in a *READ_EPR* request.
- The end of a consultation at a Point_of_service – usually resulting in an *UPDATE_EPR* request.

3.4.1 Centralised Network Model

The UK's national EPR network (Ringholm, 2007)was used as the case study for the first model. Its NPR model consists of a single server connecting to *n* facilities that in turn serve a particular population. The OMNeT++ network diagram of the model (where *n* = 2 facilities) is shown in Figure 13below:

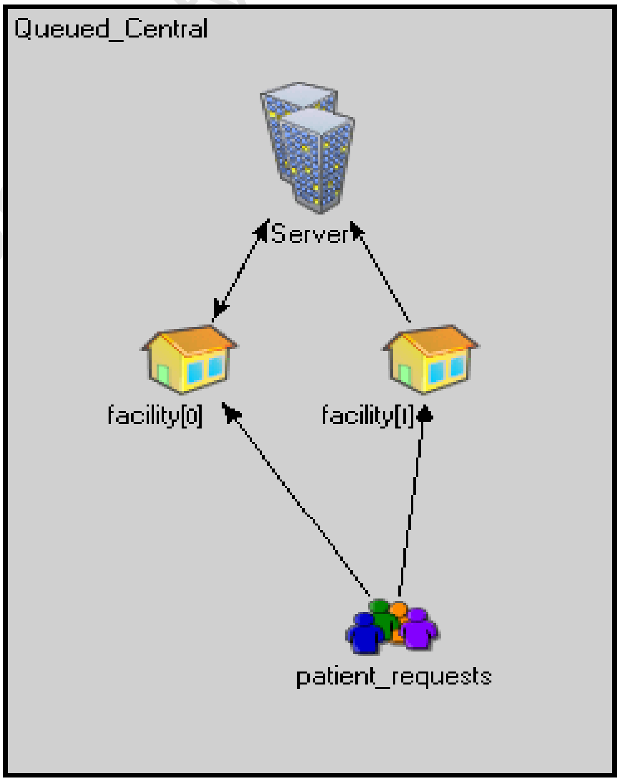


Figure 13: Centralised Network Model in OMNeT++

The centralised network has a rather consistent and simple algorithm for dealing with the two stimulating events mentioned above. Both events ultimately cause a simple request-response transaction between the facility and the server, as illustrated by the sequence diagram below:

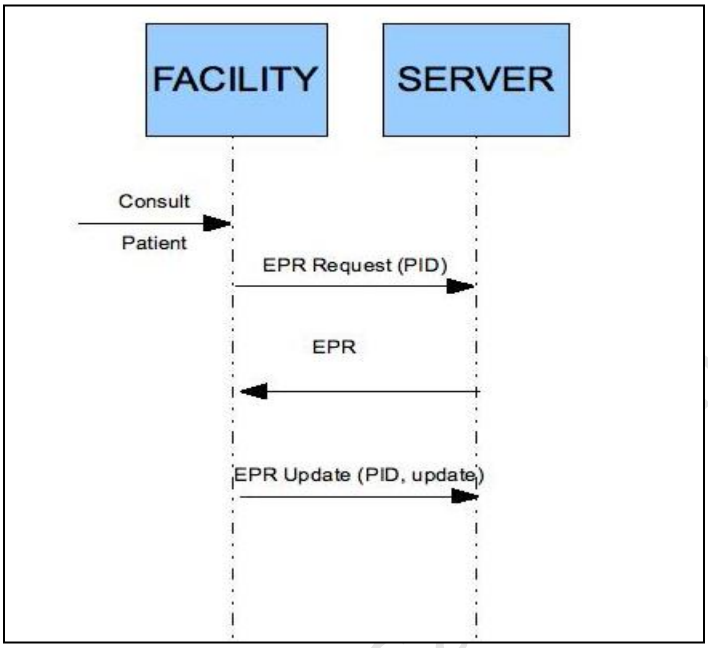


Figure 14: Centralised Network Algorithm Sequence Diagram

This network requires the following parameters:

Number of facilities – a parameter to dynamically create the number of facilities in the network.

3.4.2 Peer-to-Peer Network Model with Super Nodes

The US's proposed national EPR network (the NHIN) (Gartner, 2007) was used as the case study for the second NPR model. As discussed earlier, its architecture consists of *n* super-peer servers linked peer-to-peer, with each super-peer acting as a gateway server for a particular number of local facilities. This is represented in the OMNeT++ tool in the Figure below:

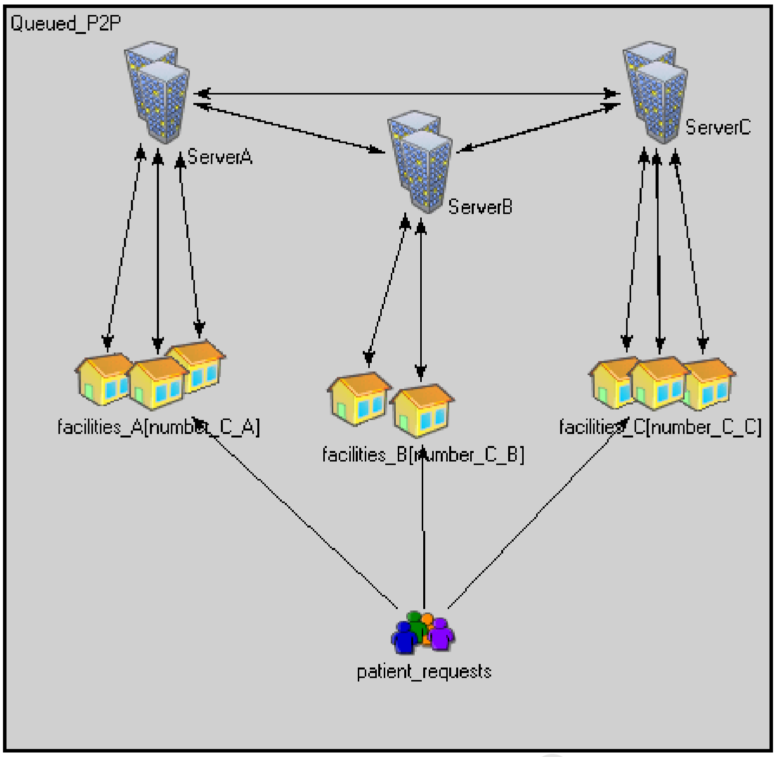


Figure 15: Peer-to-Peer Network with Super Nodes in OMNeT++

The peer-to-peer network has a much more complex network and algorithm than the previous example. The event of a patient's arrival does not always stimulate the same results in the network, and the *UPDATE_EPR* action is merely a notification to the supernodes, rather than a publication of the data to the network. These transactions between the facility and the server are illustrated by the sequence diagram below:

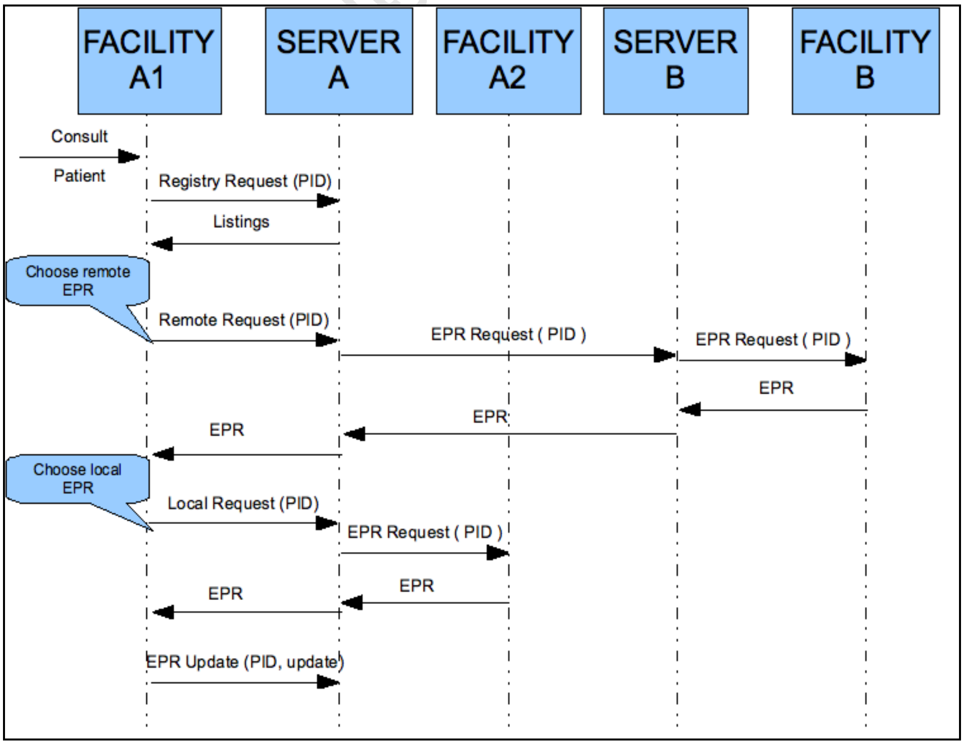


Figure 16: Peer-to-Peer Network with Super Nodes Sequence Diagram

The network size can be configured using parameters to specify the number of Super Nodes (gateways) as well as the numbers of facilities each Super Node supports. For example to produce the network shown in Figure 15above requires the following parameters:

1. Number of Super Node = 3
2. Number of facilities connected to server A = 2
3. Number of facilities connected to server B = 2
4. Number of facilities connected to server C = 2

3.5 Evaluation Metrics

Now that the model and algorithms for the networks have been defined, methods of evaluation can be determined. To do this 'Network Metrics' are created that can give comparable evaluations of the different networks or different simulation runs. A Network Metric is a calculation that gathers data as the simulation runs and uses this to provide a quantitative measurement at the end of the simulation.

Predominantly, network simulations compute metrics at the network levels such as bandwidth and error rate. For this research this was broadened to include two additional groups: application metrics and service metrics – with applications being the software or algorithm being used in the network and services being the real world work being supported. Interestingly, these groups of networks relate to each other in a similar way to that of the network stack. These are depicted in the diagram below:

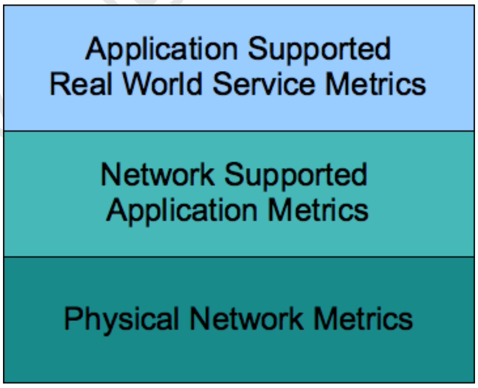


Figure 17: Network Metric Stack of Dependencies

The stack of metrics is used to illustrate the dependency of real world services on the layers below. Poor performance of the physical network could reasonably be assumed to affect the performance (to some degree) of the layers above it.

3.5.1 Network Metrics

Two sub-categories of network metrics were also distinguished: performance and demand.

Performance metrics include*:*

- **Round Trip Time**
  This is the observed time elapsed by a facility waiting for a successful response to a EPR read request. It is calculated by recording the difference between the time the request was sent and the time at which a successful response was received. To implement this in OMNeT++ the request time was encoded into the request packet, and the server copied this value into the response. Then, when the facility receives a successful response, it performs the calculation (using the simulation time) and records the answer in a *OM_vector* data structure. The OM_vector then allows calculation of the minimum, average and maximum values collected.
- **Node Availability**
  This is the percentage of requests to a node (usually the server) that receive a successful response. Obviously this is a 'perceived' value of the node in question, and thus a particular server can have different Availability values from different clients. In OMNeT++, this is recorded as the number of successful requests divided by the total number of requests.
- **Transmission Errors**
  This is the number of packets that are unsuccessfully transferred from a node to its destination. This is recorded as the number of unsuccessful requests.

Demand metrics include*:*

- **Bandwidth Used**
  This is the total amount of data transferred or received by the node for the whole duration of the simulation. To calculate this in OMNeT++, a variable is used to count each packet sent and received. In future, when packets may be modelled to have different payload sizes, this would be extended to record the actual size of the packet.
- **Average Bit Rate (Bandwidth Demand)**
  Commonly known as bit rate, this metric represents the amount of data transferred in a particular window and ultimately collects the trend across the whole simulation. As above an *OM_vector* is used and can output the minimum, maximum and average of the recorded values. The graph of this metric can show at which points in the simulation the bandwidth demand was the greatest.

3.5.2 Application Metrics

Application metrics measure the reliability and performance of the network at the application or algorithm level. These are intended to give measurements that can aid in understanding how a a particular application performs on a particular network and under certain conditions. As illustrated above, these metrics have a correlation with the lower-level network metrics, meaning that poor network performance usually will cause implications at the application level.

- **Record Freshness Metric**
  This is a measure of how up-to-date a particular patient record is. It is calculated as the number of differences between the local copy and the most up to date version of the EPR that is being used. Some of the literature refers to this particular method of calculating data freshness as obsolescence (Bouzeghoub, 2004). To implement this in OMNeT++, EPRs were recorded in a RecordRepository data structure (as described earlier) at each of the nodes that dealt with patients. There was also a control RecordRepository in the Population node – which is always up to date as all patient visits were generated there. While a patient visit was conducted, the difference between the local and control copy of the EPR could be calculated and recorded.

3.5.3 Service Metrics

Service metrics measure the impact of a network and application's performance on the *real-world-services* that they are intended to support. An example of one such metric is described below:

- **Facility's Queue Length**
  This metric monitors the length of the patient queue(s) in a facility. It is assumed that any network or algorithm delays could also cause knock-on effects to the facilities' processing of arriving patients. The metric is calculated by recording the length of the queue and the time every time its state changes. In OMNeT++ the *OM_vector* data structure can again be used to collect the numbers and out put the minimum, maximum and average length of the queue. Also, plotting the length of the queue against the time of the simulation can potentially reveal some informative trends.

3.6 Summary

This chapter has demonstrated how real world examples were used to create a conceptual model to represent different NPR solutions. The model consists of Facilities, Servers, Connections as well as other data structures needed to represent the application behaviour of the NPR. The model also used a module to represent the local population consisting of four patient types, with unique patterns of facility usage. Other contextual challenges, such as message loss or facility downtime were also included.

Finally, the evaluation metrics illustrated the need to evaluate the solution at different levels - at the network level but also at the application an service level.

The following chapter will look at how the conceptual and programmed model were validated and verified. It will consider what scope of application the model has and what the resulting confidence of using the model may be.

Chapter

4 Validation, Verification & Testing

4.1 Introduction

The NPR network model described in Chapter 4 was designed for the modelling of many potential NPR solutions to evaluate their performance in a given context. Before using it to evaluate potential NPR solutions, the accuracy and reliability of the model to perform simulation studies has to be determined.

This chapter investigates the quality and correctness of the presented solution, in order to determine the scope and limitations of its potential use as an evaluation tool.

Specifically, the goals of this chapter can be stipulated by three key questions:

1. *How valid and complete is the conceptual model's representation of the real world?*
2. *How accurately can the programmed simulation model represent real world behaviours?*
3. *What are the implications of the answers to the preceding questions on the use of the programmed model as an evaluation tool?*

Predominantly in simulation studies, a set of techniques known as Model Validation, Verification and Testing (VVT) as introduced by Balci (Balci, VVT Life Cycle, 1994) are used to find the relative correctness of the model and thus its usefulness and applicability. This chapter will continue to introduce the VVT rationale, examine its use on the model and ultimately use these outcomes to attempt to answer the questions above.

4.2 Background, Scope & Limitations of VVT

VVT Definition
Any model designed for a simulation study requires an assessment of its correctness and applicability before it can be used in practice. Establishing a level of confidence in the simulation’s results and understanding its limitations are critical to its effective use. Ultimately, this is an effort to contribute to the credibility of the overall simulation study the model is used in.

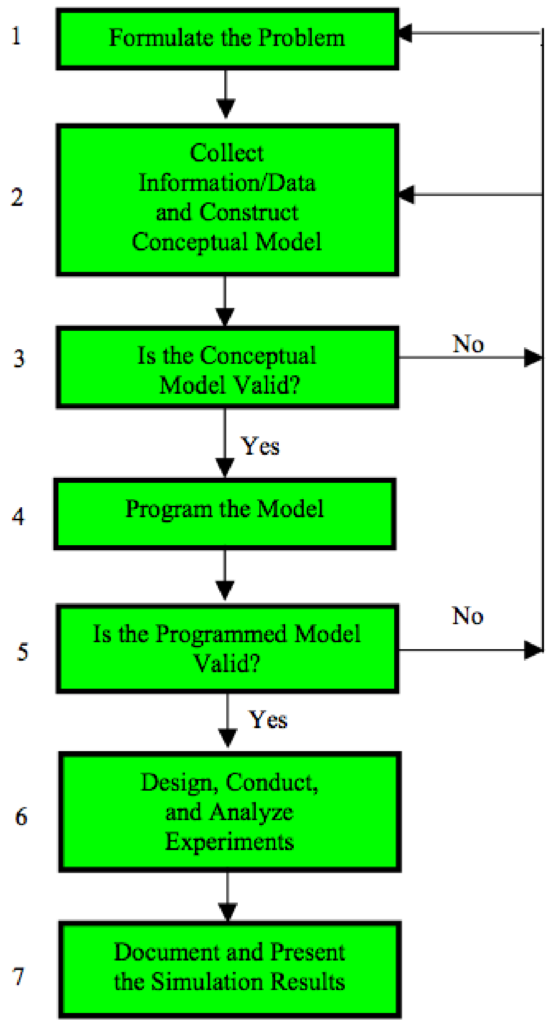


Figure 18: Law's Seven-Step Approach for Conducting a Simulation Study

In Figure 18, Law (Law, Designing a simulation study, 2003) defined the "seven-step approach for conducting a successful simulation study", and clearly highlighted two points to assess the model. Firstly, after constructing the conceptual model of a real world system, the next step is to assess this model's validity. Similarly, once the model is represented in some simulation program, the programmed model is also tested for validity. Balci (Balci, Simulation Models, 1995) develops this approach by distinguishing between the two tasks. Thus, model validation is the process of establishing whether the model is a correct representation of the relevant pieces of the real world system, while model verification deals with the accuracy to which the programmed model replicates real world behaviour.

Balci extrapolates from this by defining the concept of Validation, Verification and Testing (VVT) for simulation studies (Balci, VVT Life Cycle, 1994). VVT has since widely been referenced as a methodology to establish the accuracy and credibility of modelling and simulation research. Testing is added to the definitions above as the practice of performing various experiments on the model with specific test data, in order to yield results to help in understanding the model's behaviour. This makes testing an integral part of VVT as the outputs of planned testing can be used for both validation and verification tasks.

Scope and Limitations

No model can be as complete and as accurate as the real world itself. This means that an acceptable level of credibility needs to be determined for the simulation study's outputs. In practice, the study objectives will dictate what level of accuracy or confidence is acceptable for the model to be deemed satisfactory. This implies that model validation is strongly tied to the accurate definition of the defined research problem. Thus, model validation partly consists of checking that all relevant parts of the real-world system are adequately represented, as well as ensuring that the study objectives are adequately defined and can be measured by the outputs of the model. This is often referred to as model completeness or model coverage (Law, Building valid models, 2001).

As simulation models are usually built on areas outside the average programmer's knowledge, Subject Matter Experts (SMEs) of the modelled domain are often involved in the validation process (Law, Building valid models, 2001). Their subjective assessment of the model is known as Independent Verification & Validation (IV&V) and is often used to evaluate the model completeness and acceptable confidence concepts that are described above. Techniques such as animation and program traces are used to engage with the SMEs and allow them to assess to what degree the program is replicating the behaviour of the real world.

However, unless the SME is also the sponsor of the research, engagement can often greatly increase the costs of the study. Moreover, Sargent (Sargent, Model verification, 2007) shows that efforts to increase model confidence actually lead to an *exponential* increase in cost, as illustrated in Figure 19. This has further implications for the design and validation of models, as the level of desired (and feasible) confidence is often determined by the project budget.

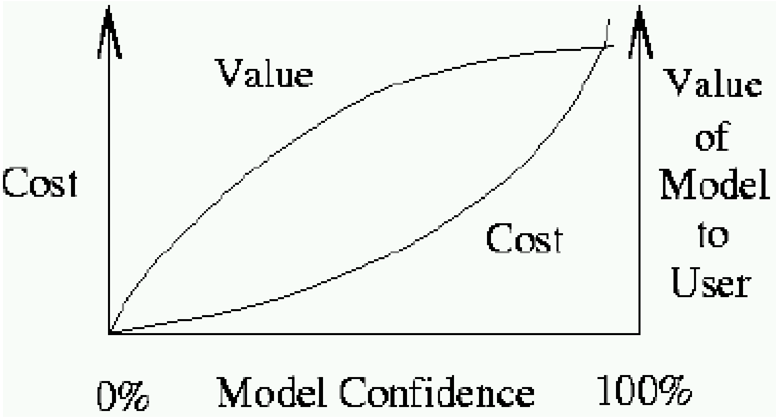


Figure 19: Sargent's Model Confidence Relationship to Cost

Beyond SMEs and IV&V, there are several other validation techniques, including event tracing, data flow analysis and historical data black box tests(Sargent, Verification and Validation of simulation models, 1998). Importantly, many of these techniques rely on the availability of similar models, experiments or data from the real world systems under investigation. This can lead to somewhat of a paradox as stated by Law (Law, Building valid models, 2001):

> “Validation is inherently difficult because the systems under study are either proposed modifications of an existing system or represent completely new systems. Thus, it is difficult, if not impossible, to obtain performance measures from the actual systems of interest to use for validation purposes”

Model verification tests and techniques are perhaps easier to define. They focus on the accuracy of the model’s ability to represent real world behaviour. Extreme condition tests, boundary tests, parameter resistivity analysis and other more quantitative experiments can be performed to analyse the programmed model’s accuracy. Compared to validation, verification provides more techniques to assess the aspects of the programmed model – but many tests can still be restricted by the same barriers of real world reference data referred to earlier. Furthermore, while validation can be quite dependant on the subjective opinion of SMEs, verification is most valuable when it can be compared to the outputs of existing systems.

As mentioned earlier, testing is crucial in providing outputs for the validation and verification of the model. Several authors underline the need for this testing (and in fact the whole of VVT) to be performed throughout the entire life cycle of a simulation study. This means that the conceptual model will often evolve over the iterations of the study, as every improvement in the model feeds insights and corrections into the next iteration of design. Balci’s 7th Principle of Simulation VV&T (Balci, Simulation Models, 1995) tells us complete or exhaustive testing of a model is not typically possible, as this would require testing under *all* possible input parameters and model configurations. Thus, testing is focused on tests that deliver the most benefit to the study objectives. Credibility can obviously still be improved by the percentage of valid inputs covered.

While VVT has now been defined, the limitations of its outputs need to be considered. The original purpose of VVT was to establish the accuracy of the model for representing real-world behaviour. To consider this, some fundamentals of VVT need to be considered:

1. As mentioned above, Balci’s 7th Principle of Simulation VV&T states that exhaustive testing across all possible inputs is not possible, and that tests should be focused on the study objectives.
2. Importantly, this has implications for Balci’s 6th Principle of Simulation VV&T, which states that the prescribed conditions for which the model’s credibility was established are the *only* conditions for which its credibility can be claimed.
3. Similarly, in terms of model validity, Law (Law, Building valid models, 2001) cautions that “*There is no such thing as absolute model validity, nor is it even desired. Indeed, a model is supposed to be an abstraction and simplification of reality”.* This also contains implications that real world systems are predominantly too vast to be modelled in totality, and that, rather, the specific area of problem interest should be focused on.

4. As such, Balci's 2nd Principle of Simulation VV&T states that the outcome of model VVT should be considered as a degree of credibility on a scale of 0 to 100, as opposed to the binary option of correct or incorrect.
5. Balci's 12th Principle of Simulation VV&T also cautions that "Successfully testing each sub-model does not imply overall model credibility".

In summary, these fundamentals of the VVT of simulation studies show that the decision of what accuracy is acceptable is a critical part of defining the study objectives.The rest of the chapter will now document the validation, verification and testing performed on the NPR model and compare these results to the study objectives.

4.3 Conceptual Model Validity

To recall, the goal of model validity is to define: *how valid and complete is the conceptual model's representation of the real world?* Considering the limitations and challenges to validation listed in the background, such as the lack of real world system data, the aim of the validation experiments has to be determined. As described above, a model can seldom be completely correct or incorrect and, as such, it is prudent to determine the degree of validity that is acceptable. The conceptual validity question above can be combined with the main research question ("*Is simulation a feasible technique to evaluate designs for a National Patient Record in South Africa*") to produce the following validity targets:

- The model can represent different NPR electronic networks,
- The model can represent the SA population and contextual challenges,
- The model can provide operational metrics for the networks.

In considering the techniques for validation, one particular challenge in proving the validity of this particular model is the lack of available system data from similar systems. Furthermore, as pointed out by Sargent (Sargent, Verification and Validation of simulation models, 1998), IV&V as a validation approach is often used when the development of the simulation study has a large cost and thus the increased costs of involving the SMEs is justified. For these reasons, the validation effort focuses on more introspective techniques that look at the composition of the model compared to the available literature on the subject of NPRs.

4.3.1 Requirements Benchmarking

The first step in validating the NPR model is to validate it against the functional requirements that were established for the model earlier. As listed in Table 1 below, these functional requirements described the initial real world actions, processes and systems that were defined to be the key parts of an NPR system.

Table 8: Functional Requirements for an NPR

Func. Req.	Requirements for NPR simulation Model	Real World Reference
FR01	Patients are able to visit all facilities over a period of time.	SA utilisation rates
FR02	Facilities to update patient visit in individual's EPR	US / UK system specifications
FR03	Individual's EPR available for all facilities to access	US / UK system specifications
FR04	Connections susceptible to lost packets	Health Systems Trust, SA Gov
FR05	Nodes susceptible to periods of downtime	Eskom
FR06	Calculate Network metrics: Availability, RTT, BW Cap, Bandwidth rates	
FR07	Calculate Application Metrics: Data Freshness	

Where possible, the requirements can be compared experimentally with real world data to understand the model's ability to represent real world behaviour. The next section describes the experiments and their results.

4.3.2 Validation Tests

Experiment 1
Functional Requirement - FR01 Patient's facility utilisation
Aim – To establish whether the population's provision of patient visits to facilities is realistic.
Method – To establish a comparable benchmark, the volume and frequency of patient visits in the model are monitored and then compared with existing statistics of healthcare facility usage in South Africa (Health Systems Trust, 2004).
Results – As shown inTable 9, the ideal and actual average facility utilisation rates (the average number of visits to a Primary Health Care centre per capita per year) are in the order of 2 to 3 visits per year per patient. Table 10 shows the results of four experiments using the NPR model using a fixed population and two different amounts of facilities in each simulation.

Table 9: South African benchmarks for facility utilisation rates

Government modelled ideal Utilisation rate	Actual average SA Utilisation rate
3	2.3

Table 10: Validation Test 1 on Utilisation

Experiment Id	Network	Population	Facilities	Time (days)	Total recorded Visits across Facilities	Utilisation Rate NPR Model
R212	US	100 000	3	365	274636	2.74
R213	UK	100 000	3	365	309257	3.09
R218	US	100 000	20	365	278448	2.78
R219	UK	100 000	20	365	421696	4.22

InTable 10, the results shown in the ‘Utilisation Rate NPR Model’ column illustrate that although the model’s values do tend to be slightly higher than the given benchmarks, they are still well within one standard deviation. Based on the specific needs of a simulation study, this could be considered reasonably representative of the real world behaviour.

The population model component recognises that it is the unique identity of the patient that determines its facility affinity and facility utilisation. The modelling of patient types (acute, chronic and random), which influenced both the visit frequency and the facility affinity, appears to have helped to keep the utilisation rate within the correct order of magnitude.

Patient type is only one of many variables that would contribute to facility utilisation, and new model components such as facility type and some measure of facility-patient proximity could potentially contribute to the modelling of facility affinity for each patient. This indeed could improve the accuracy of the overall model as well as extend its flexibility to model different types of populations and patients.

Experiment 2
Functional Requirement - FR02 - Facilities to update patient visit in individual’s EPR and FR03 - Individual’s EPR available for all facilities to access.
Aim – To validate that many facilities are able to read and maintain a shared record for a single patient.
Method – As suggested by several authors (Law, Sargent, Balci), system traces can be used to validate that the simulation is performing the desired real world function. For this, the simplest trace was used – that of identifying one patient in a simulation – and checked that the patient’s EPR copy at the different facilities reflected their visits. The simplest way to do this is to compare the local copies to the master copy of the EPR (the perfect copy that was used to monitor data freshness). To ensure that the modelled challenges, such as downtime and connection failures, did not interfere with the test, the network conditions were set to perfect – with no loss or downtime in the network. The experiment was run with the parameters shown in Table 11 below:

Table 11: Experiment 2 Input Parameters

Exp		Facilities				Population	Server	Simulation
#	NW	#	Loss prob	DT Prob	Av. DT period	Pop size	Loss prob	Duration
2.1	US	3	0	0	0	1	0	1 year
2.2	UK	3	0	0	0	1	0	1 year

Results – For the US NPR model, Table 12below shows the sequence of the single patient’s five visits to the three facilities in the simulation, listed sequentially in the ‘Event’ column. The EPR record traces indicate the local copy of the EPR for the patient after each event, in comparison with the copy of the Master Record of the EPR, which records all visits independently.

Table 12: Validation Test 2 on EPR Update Traces US Model

Event		EPR Record Trace Output After Event			
#	**Description**	**Master Record**	**Facility 1**	**Facility 2**	**Facility 3**
1	Patient Visits Facility 2	<Visit 1, Facility 2>		<Visit 1, Facility 2>	
2	Patient Visits Facility 2	<Visit 1, Facility 2> <Visit 2, Facility 2>		<Visit 1, Facility 2> <Visit 2, Facility 2>	
3	Patient Visits Facility 1	<Visit 1, Facility 2> <Visit 2, Facility 2> <Visit 3, Facility 1>	<Visit 1, Facility 2> <Visit 2, Facility 2> <Visit 3, Facility 1>	<Visit 1, Facility 2> <Visit 2, Facility 2>	
4	Patient Visits Facility 2	<Visit 1, Facility 2> <Visit 2, Facility 2> <Visit 3, Facility 1> <Visit 4, Facility2>	<Visit 1, Facility 2> <Visit 2, Facility 2> <Visit 3, Facility 1>	<Visit 1, Facility 2> <Visit 2, Facility 2> <Visit 3, Facility 1> <Visit 4, Facility2>	
5	Patient Visits Facility 3	<Visit 1, Facility 2> <Visit 2, Facility 2> <Visit 3, Facility 1> <Visit 4, Facility2> <Visit 5, Facility3>	<Visit 1, Facility 2> <Visit 2, Facility 2> <Visit 3, Facility 1>	<Visit 1, Facility 2> <Visit 2, Facility 2> <Visit 3, Facility 1> <Visit 4, Facility2>	<Visit 1, Facility 2> <Visit 2, Facility 2> <Visit 3, Facility 1> <Visit 4, Facility2> <Visit 5, Facility3>

The traces show that the local EPRs and Master Record are correctly maintained in a manner consistent with the US algorithm (see appendix B – NHIN prototype).

For the UK model, under perfect conditions, the consistent maintenance of the patient's EPR is less complex as every facility synchronises with the central server on every visit. Thus the algorithm's accuracy really depends on the accurate maintenance of the patient record being kept at the server, in relation to the Master Record.

Table 13: Validation Test 2.2 on EPR Update Traces UK Model

#	Description	Master Record	Server Record
1	Patient Visits Facility 2	<Visit 1, Facility 2>	<Visit 1, Facility 2>
2	Patient Visits Facility 2	<Visit 1, Facility 2> <Visit 2, Facility 2>	<Visit 1, Facility 2> <Visit 2, Facility 2>
3	Patient Visits Facility 1	<Visit 1, Facility 2> <Visit 2, Facility 2> <Visit 3, Facility 1>	<Visit 1, Facility 2> <Visit 2, Facility 2> <Visit 3, Facility 1>
4	Patient Visits Facility 2	<Visit 1, Facility 2> <Visit 2, Facility 2> <Visit 3, Facility 1> <Visit 4, Facility2>	<Visit 1, Facility 2> <Visit 2, Facility 2> <Visit 3, Facility 1> <Visit 4, Facility2>
5	Patient Visits Facility 3	<Visit 1, Facility 2> <Visit 2, Facility 2> <Visit 3, Facility 1> <Visit 4, Facility2> <Visit 5, Facility3>	<Visit 1, Facility 2> <Visit 2, Facility 2> <Visit 3, Facility 1> <Visit 4, Facility2> <Visit 5, Facility3>

Table 13shows a similar trace to Table 12but of the records maintained at the Server and the Master Record. It illustrates that under perfect network conditions, the record copy at the server is indeed correctly maintained for all facility visits.

Experiment 3

Functional Requirement - FR04 & FR05 Simulate a Facility's unstable connection to the network.

Aim – To establish how realistic the modelling of downtime is in the NPR model.

Method – To see how the modelled downtime and connectivity represents the real world, this experiment compares existing real world data to the simulated downtime pattern for a facility in the simulation. In contrast to the previous experiments, the downtime simulation of facilities was independent of the network (US or UK) implemented, and therefore only one simulation run was needed. The simulation chosen to most closely represent a facility in the developing context was with the following input parameters:

Table 14: Experiment 3 Input Parameters

Exp		Facilities					Population	Simulation
#	NW	#	Loss prob	DT Prob	Av. DT period	Daily Visit Window	Pop size	Duration
220	US	3	0.001	0.01	600	8 hours	100000	1 year
320	US	3	0.001	0.001	600	8 hours	100000	1 year

Results

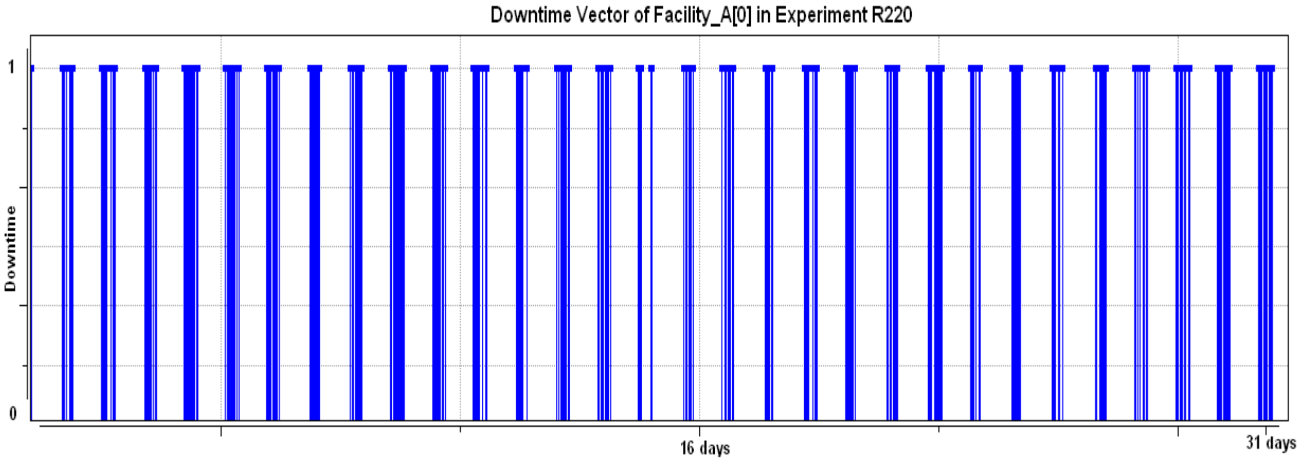


Figure 20: Downtime at Facility over a month's period

Figure 20shows the downtime pattern of a facility over a period of 30 days. Downtime (lack of connectivity to the network) is illustrated by a value of 1 for downtime and 0 for the time the node was connected to the network. In the graph a distinct recurring pattern is evident - this is caused by the 'daily visit window' parameter, which models the number of hours a facility is open each day. In this experiment this value is 8 hours, while for the remaining 16 hours of the day downtime is not monitored – this explains the repeated pattern and also indicates that downtime is a daily occurrence.

Similarly, Figure 21below illustrates the downtime pattern over a single day for the same facility. This shows that the periods of downtime are indeed random within a given day and shows 11 downtime periods, averaging 10 minutes in duration. For this particular day it means that, of the 8 hours the facility was open, it was unable to connect to the network for 1 hour 50 minutes (cumulative) or approximately 23% of its time.

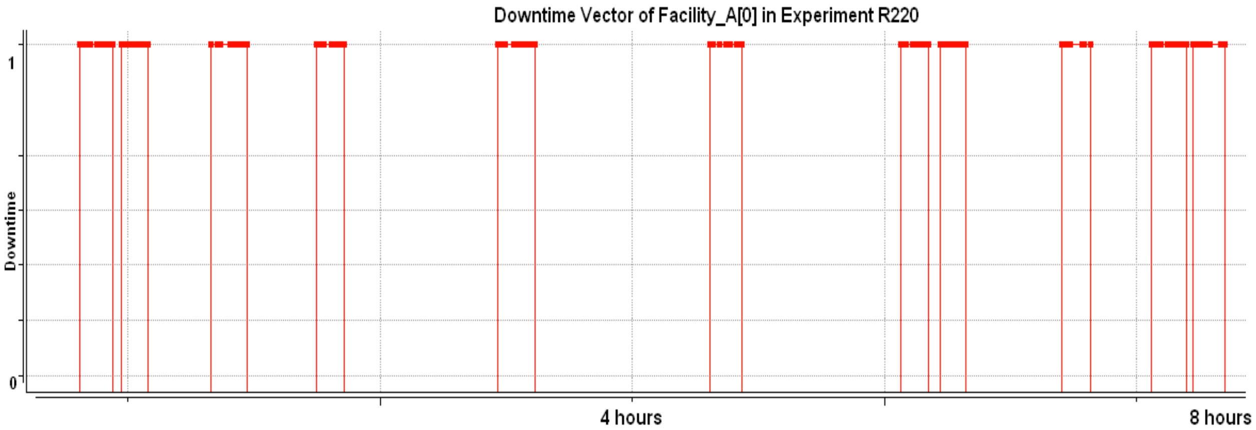


Figure 21: Downtime at a Facility over the period of 8 hours

The real world data on electricity outages in areas of South Africa (National Energy Regulator of South Africa, 2008) states that over the periods of early 2006 to late 2007, electricity demand was short of supply in the range of 6% to 22%. The ideal target was a 10% average shortage, but realistically this was closer to a 14% average. This is a broad statistic of the wider South African environment and does not take into consideration local challenges and barriers (such as cable theft) in the downtime statistics. Using this to form the assumption that rural primary health care facilities would suffer *above average* downtime due to local challenges, the outcomes of the modelled PHC facilities appear more reasonable.

Furthermore, this part of the model also shows a strong correlation to the input parameters and as more is understood about downtime patterns, the model can be adjusted and customised using these parameters. To illustrate, the additional experiment (320) listed in Table 14 has the exact same parameters as the preceding experiment except the downtime probability, which is reduced by a tenth (1% to 0.1%). As shown in *Figure 22* below, this reduces the number of downtime events over a month considerably, and reduces the average downtime to approximately 0.5% of the time the facilities are open.

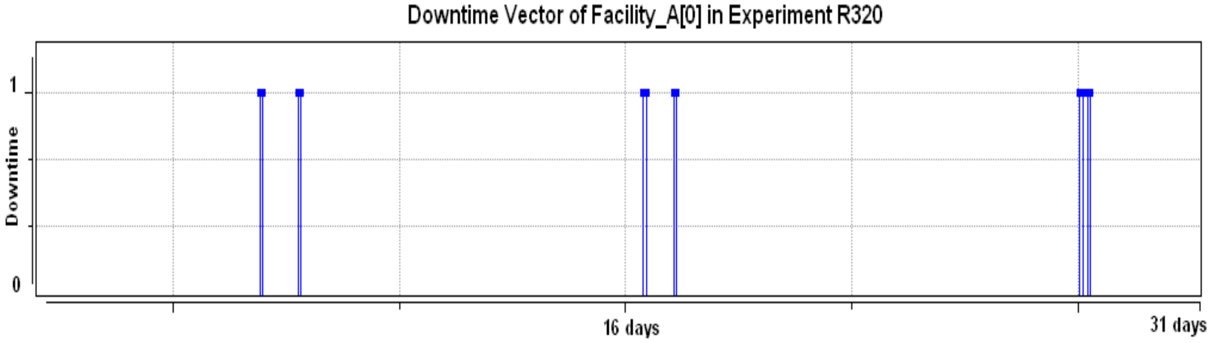


Figure 22: Experiment 320: Downtime at Facility over a month's period

While the amount of downtime simulated appears to be reasonably acceptable, a question remains over the *patterns* of downtime. The daily occurrence of downtime, as illustrated in the first simulation (Figure 20), is yet to be validated against other possibilities of more extreme periods of downtime punctuating a more consistent connection to the network. This alternative could certainly provide the same average downtime, but have quite different impacts on operations at a daily level. The lack of event level sample data from real facilities hindered this step in validation, and the gathering of such data could warrant a dedicated research effort of its own. Once such data is available, it could be used to test the boundaries of the model's ability to represent different conditions.

4.4 Programmed Model Verification

As discussed earlier, Model Verification is the process of establishing a confidence rating of the programmed model's ability to represent *the workings and behaviour* of the real world system. Again, where Model Validation aims to test if the right model was created, Model Verification is testing if the model is created right. These techniques tend to be more quantitative in nature as definite processes, behaviours and actions can be numerically measured.

This section first introduces the concept of unit tests and how they formed a critical part of the verification process, by breaking the programmed model into sub-models that could be tested and verified independently.

4.4.1 Unit Tests

Experiment 4
Definition – Automated and heuristic tests to maintain the accuracy of various sub-model components and algorithms.
Aim – Determine if the programmed algorithms and components of the NPR model replicate the desired real world behaviour.
Method – Borrowed from extreme programming, the concept of Unit Tests and *test-driven development* (Beck, 2004) was used to test subsets of the programmed model. As new components and behaviours were added to the system, a test case would be designed and implemented in a simulation run. Critical to the design was an expected output based on the controlled input parameters. Ultimately, when the unit test is run it is testing to see if the

functionality still transforms the given inputs to the expected outputs (this could be considered an automated sub-model black box test).

Building the unit tests in this method helped ensure that the core functionality of the model maintained its integrity as the model evolved. When new model components are added, the “Unit Test” simulations can be re-run to ensure they have not been inadvertently broken. As such, the list of unit test develops along with the model and aligns itself well to the iterative design process, as the results of a failed test may shed insight on the latest changes to the model.

Table 15 below lists the unit tests that were used for both the UK and US NPR models. Each unit test tests a model component or behaviour using a controlled set of inputs. Each test also has some success criteria – the desired outcome of a successful test. The ‘verification Tool’ column describes how the outcomes of the test are monitored, using many of the monitoring tools of OMNeT++.

Table 15: List of Unit Tests used in Verification

Test	Unit Test	Controlled Inputs	Verification Tool	Success Criteria
1	Population can generate random and distributed patient visits	Population = 3 Facilities = 3	Simulation animation and Event Trace	All patients are observed to visit at reasonably random times and facilities.
2	Facilities can queue patient arrivals into EPR requests	Population = 3 Facilities = 1	Simulation animation and Event Trace	All patients are observed to be added sequentially to the Facility’s queue and then in turn cause an EPR update at the Point-of-Service.
3	Correctly update the local EPR	Population = 1 Facilities = 3	Event Trace, Data Structure Watch and System outputs	Each of the patients’ visits are correctly recorded in the local patient record.
4	Correctly update the Server visit logs	Population = 1 Facilities = 3	Event Trace, Data Structure Watch and System outputs	Each of the patients visits are correctly recorded in the server’s visit logs and patient record.
5	Correctly route an EPR request	Population = 1 Facilities = 3	Simulation animation and Event Trace	A remote patient request is correctly gathered, retrieved and delivered.
6	Correctly propagate an EPR update	Population = 1 Facilities = 1	Event Trace, Data Structure Watch and System outputs	After the patient visit, the update is correctly recorded on the server.
7	Correctly handle timeout	Population = 1 Facilities = 1 Server connection loss probability = 1.0	Simulation animation and Event Trace	The initial messages sent to the server should be lost, and the facilities should try to resend a finite number of times.
8	Correctly calculate bandwidth	Population = 1 Facilities = 1 All connection loss probability = 0 All downtime probability = 0	Event Trace, Data Structure Watch and System outputs	For the fixed number of visits given in the population’s Master Record, calculate the expected bandwidth for the facility and server, then check if this matches the measured values.
9	Correctly calculate data freshness	Population = 1 Facilities = 3 connection loss probability = 1.0	Event Trace, Data Structure Watch and System outputs	Compare the monitored values of Data Freshness at each facility to the expected values after each event.
10	Correctly simulate Downtime of Facilities	Population = 1 Facilities = 3 connection loss probability = 1.0	Simulation Animation, Event Trace and Data Structure Watch	Watching the patient visits to facilities, observe the ‘uptime’ vector of each facility and monitor if it correctly handles messages.

Results
These unit tests were evolved over the various design iterations in the study, and the model implementation was corrected until all passed at each stage. Thus, with the final NPR model design, all of the above unit tests passed their success criteria.

4.4.2 Seed Independence Testing

Experiment 5
Definition – Most simulations depend on some form of random input and random number generators (RNG) are used to produce these number series to be fed into the model. For a given simulation, different series of random numbers should not significantly affect the simulation outputs. Effectively, the model should be robust enough for its overall outcomes to be independent of the random series, despite the variation in individual values.
Aim – To determine if the NPR model's results are dependant on the particular RNG seed used.
Method – Different RNG series are generated from different seed values, so this experiment will evaluate the differences in model output across different seed values. To do this a standard experiment was established (1000 patient population, 3 Facilities) and run with two different seed values. As an extra precaution, the experiment was performed twice with the first seed, to ensure that results were consistent given the same seed. Again, the experiments were run on both the US and UK models.
Results

Table 16a: Seed Independence Testing of UK Model

RNG	Metrics at Facility			Metrics at Server			Seed Parameter
1		Avg	Max		Avg	Max	seed-0 = 3454326
	Periodic Bandwidth	5	18	Periodic Bandwidth	10	33	
	Data Freshness	0	5	Total Requests	42983		
1		Avg	Max		Avg	Max	seed-0 = 3454326
	Periodic Bandwidth	5	18	Periodic Bandwidth	10	33	
	Data Freshness	0	5	Total Requests	42983		
2		Avg	Max		Avg	Max	seed-1 = 2794537
	Periodic Bandwidth	5	21	Periodic Bandwidth	10	33	
	Data Freshness	0	5	Total Requests	43167		

Table 16b: Seed Independence Testing of US Model

RNG	Metrics at Facility			Metrics at Server			Seed Parameter
1		Avg	Max		Avg	Max	seed-0 = 3454326
	Periodic Bandwidth	8	51	Periodic Bandwidth	9.8	55	
	Data Freshness	1	6	Total Requests	41318		
1		Avg	Max		Avg	Max	seed-0 = 3454326
	Periodic Bandwidth	8	51	Periodic Bandwidth	9.8	55	
	Data Freshness	1	6	Total Requests	41318		
2		Avg	Max		Avg	Max	seed-1 = 2794537
	Periodic Bandwidth	8	36	Periodic Bandwidth	9.8	54	
	Data Freshness	1	7	Total Requests	41730		

Table 16a and 16b show the results from the UK and US experiments respectively. In both sets of results there are two clear findings: firstly, that given the same seed the simulation provides identical outputs; and secondly, that given a different seed, the difference in randomisation only affects the outliers (maximum values of bandwidth and data freshness) but not the overall averages.

4.4.3 Parameter Sensitivity Testing

Experiment 6
Definition –Sensitivity analysis is testing which model input parameters have the greatest effect on programmed model behaviour (Law, Building valid models, 2001).
Aim – To determine which model input parameters have the greatest effect on programmed model behaviour.
Method – Simple simulations with default parameter values for both NPR models were run as benchmarks for the default behaviour of the network. Then each of the parameters was increased by an order of magnitude in a dedicated simulation and the impact of the outcomes recorded. Comparisons between each of the parameter simulations and the benchmark run were then performed to assess the role of the particular parameter. Again, the experiments were run on both the US and UK models to understand how the major parameters affect each model.

Results

Table 17: Parameter Sensitivity Results

Simulation	Component Focus	Inputs	Output					
Benchmark Run	All	Population = 10k Facilities = 3 Facility Loss Prob. = 0.005 Time = 1 year		BW Av	BW Max	BW Usage	DF Av	DF Max
			UK F	5	20	43138	0	4
			UK S	10	36	129220	-	-
			US F	10	67	57000	1.1	7
			US S	10	72	115326	-	-
Population – increase by 10^1	Population	As benchmark + Population = 100k		BW Av	BW Max	BW Usage	DF Av	DF Max
			UK F	25.5	66	311360	1	9
			UK S	54	144	924651	-	-
			US F	30	85	104474	2.5	8
			US S	37	143	208311	-	-
Facility – increase to 20	All	As benchmark + Facilities = 20		BW Av	BW Max	BW Usage	DF Av	DF Max
			UK F	1.2	9	8921	0	5
			UK S	11	32	132102	-	-
			US F	5	23	15154	1.8	8
			US S	15	76	202930	-	-
Facility Loss Probability – increase by 10^1	Facility	As benchmark + Facility Loss Prob. = 0.05		BW Av	BW Max	BW Usage	DF Av	DF Max
			UK F	4	18	28111	1.1	7
			UK S	8	33	86301	-	-
			US F	5	28	10497	2	7
			US S	6	45	21584	-	-
Simulation Time – increase by 10^1	Simulation	As benchmark + Time = 10 years		BW Av	BW Max	BW Usage	DF Av	DF Max
			UK F	5.5	24	466388	1.7	61
			UK S	10	42	467257	-	-
			US F	Incomplete				
			US S	Incomplete				

Table 17 above lists the results of the 5 simulation runs with different parameter values, and their respective outcomes. The following section analyses the parameters' impact on the bandwidth and data freshness metrics.
For bandwidth, population has a somewhat obvious effect on all bandwidth metrics and appears to be directly proportional for both models. Consistency can also be seen across bandwidth metrics in the UK model as for the same population values, the aggregation of bandwidth metrics are similar regardless of the number of facilities. Indeed the number of facilities would appear to have an inversely proportional relationship to the bandwidth metrics for a given population. In contrast, excessive loss probability and downtime has an understandably reducing effect on all bandwidth values. Similarly, an extension in the simulation time has minor effect on average bandwidth values, but an obvious increase in the cumulative BW usage (which also is linear).

Ultimately, total bandwidth is affected by the total number of visits which is affected by population size and length of the simulation duration, while average bandwidth is affected by the rate of visits, which is affected by population (and number of facilities at the facility level). In addition, all metrics can be reduced by aggressive network failure.

For data freshness metrics, an increase in population led to a noticeable increase in data freshness obsolescence values, while a change in the number of facilities produced a marginal effect on the US model and negligible effect in the UK model. Theoretically, both should be fairly independent of data freshness (certainly the UK centralised model), so it is concerning that population affects it. Furthermore, the data freshness of both models was severely affected by the extension of the simulation time – indicating that the programmed model for this appears to become unreliable over long periods (such as the 10 year period tested). These findings are concerning as the ideal hypothesis would be that only network failure would reduce the quality of data freshness – as it did in the loss probability simulation.

In summary, the number and rate of visits have the biggest impact on bandwidth metrics, while both bandwidth and data freshness metrics are subject to the network's reliability constraints. However, unreliable model behaviour has become apparent in terms of its reliability over time and the relationship between population numbers and data freshness, which needs further investigation.

4.4.4 Degeneracy Testing

Definition – Degeneracy testing is the analysis of model behaviour under extreme input parameter values.
Aim – To determine if the programmed model holds its integrity under extreme conditions and identify its range of representation.
Method – A number of experiments were run,where certain critical parameters were set to boundary values to deliberately test the execution of the model.

Results – Firstly, the population was set to 4.5million across 50 modelled facilities over the standard one year simulated time. The programmed model did not work well with extreme tests – high patient load particularly led to prolonged execution time and vast increase in memory usage. This could be a limitation of the underlying operating system, hardware or the OMNet++ simulation software, but it is more likely that there is a proportional relationship between the number of events to be simulated (in this case patient visits, downtime events etc.) and the amount of system resources required.

Further tests supported this hypothesis – in that only the overloading of parameters that produced more patient visits (i.e. simulation events) caused such a drastic increase in resource usage. To illustrate – an increase in simulated time (say 1 year to 10 years as shown in the sensitivity simulations) also increased the execution time of the simulation, as did an increase in the modelled average number of visits for a chronic patient per year (actually having an exponential effect). Thus, execution smoothness or load is proportional to the number of events, and number of events is a function of time, population size and the parameters for average visits for the different patient types.

In addition to performance problems, it also was observed in earlier sensitivity tests that the veracity of the model was questionable for time periods significantly longer than 1 year.

4.5 The Role of Testing

The sections above have demonstrated the outcomes of various testing efforts, as most of the validation and verification experiments could be classed as testing. This section highlights the methodology of testing used and how this particularly contributed to the design of the model.

Theory shows us that testing is an integral part of the development of any model. As Balci's 1st Principle of Simulation VV&T (Balci, Simulation Models, 1995) argues, testing should be incorporated into an iterative design cycle were it can inform the next stage of Model development. To illustrate the use and role of testing in this research's model development,
Table 18 below shows the timeline of model development in relation to the development of the documented tests above.

Table 18: Model & Test development Timeline

Version	Model Additions	Conceptual Tests	Unit Tests
UK 1.0	Build UK Network – Servers& Clients		
UK 1.1	Clients Send requests & receive replies from Server		
UK 1.2	Add network metric BW		Calculate BW Unit Test
Pop 1.0	Add population & patient visits		
UK 1.3	Add service Queue at the facilities		
UK 1.4	Add patient Record at the UK server		
Pop 1.1	Configure visit distribution and timing patterns	Patient Utilisation Conceptual Tests	
US 1.0	Build US Network – Servers& Clients		

US 1.1	Clients Send requests & receive replies from Server		
US 1.2	Clients Send requests & receive replies from other Clients (inc Message Routing)		Propagate requests Unit Test Route EPR request Unit Test
US 1.3	Implement Server patient lookup directories		Update visit logs Unit Test
US 1.4	Implement local Record at Facilities	EPR Updates Conceptual Tests	Update all EPR Unit Tests
US 1.4	Add downtime event for facilities		
UK 1.4	Add downtime event for facilities	Downtime Conceptual Tests	Simulate Node Downtime Unit Test
Pop 1.2	Add Master Record for Population		
US 1.5	Add Data Freshness Metric		
UK 1.5	Add Data Freshness Metric		Data Freshness Unit Test
	Both US & UK		Degeneracy Testing
	Both US & UK		Seed Independence Testing
	Both US & UK		Sensitivity Testing

Reflecting on the testing effort, some conclusions can be made on the role of testing in this research:

1. The model's design benefitted from the iterative approach, allowing its growth to be better informed.
2. The unit tests reduced the amount of time testing older features, and ensured the consistency of model performance after updates.
3. The importance and role of sensitivity and degeneracy tests could have been better planned during the study design, as it can be difficult to rectify these issues at the end of the research. Alternatively, the research and model could get potentially side-tracked if this is focused on too much, so there is some trade-off.

4.6 Results & Analysis

This chapter has shown how several established VVT techniques and methodologies were applied to the NPR model. This section analyses these results to understand the scope and limitations of the model, as well as its level of confidence.

As discussed earlier the acceptable level of credibility was specified as the following targets:

- The model can represent different NPR networks.
- The model can represent the SA population and contextual challenges.
- The model can provide operational metrics for the networks.

Considering these aims, the three initial questions of VVT can be answered.

Question 1 - How valid and complete is the conceptual model's representation of the real world?
The experiments have shown that the model can successfully replicate the basic EPR maintenance functionality of both the UK and US NPR models. It also has shown it can correctly represent the contextual challenges and behaviour of the local population and network infrastructure challenges. Furthermore, it has shown that it can calculate quantitative metrics of bandwidth and data freshness on the networks. However, some limitations on these outcomes include:

1. The model can currently only replicate the algorithms of the NPR networks discussed – further model programming would be required to include other EPR sharing algorithms.
2. The model can currently only represent local challenges of network connection delay or failure, and node failure. Other challenges, such as usability, accessibility, hardware delays and interoperability were not considered.
3. The model currently contains areas of over-detail, such as the queued patient line and points-of-service model at facilities, which did not show a critical effect on the outcomes and could have possibly been replaced with a simpler approach.
4. The model also contains areas of under-detail, such as only one facility type, downtime event generation and facility affinity algorithm which, with further research, could potentially improve the model completeness.

To illustrate this last point, great effort was put into modelling the arrival of patients – which ended up not having a large effect on the behaviour of the model, as it really is a function of the processing ability of the facility. Ultimately, the lack of more than one Point of Service could have been a major drawback in the number of facility types that could possibly have been represented. Perhaps rather the facility should have been modelled as a black box that takes input patient parameters and then generates network requests based on some customised function per facility type. This indeed may simplify the further development of the model to include new facility types such as hospitals, laboratories or pharmacies. While this modelling does conceptually reflect more detail of the patient processing process, it is perhaps an area of unnecessary detail that could be better represented by a few assumptions and an equation.

In contrast, modelling the different patient types did turn out to be highly relevant – as shown earlier, the number of visits plays a big influence on the outcomes of a simulation.

Question 2 - How accurately can the programmed simulation model represent real world behaviours?

The verification tests have shown that the algorithms and components of the networks can be reasonably replicated as per the specifications, and that network metrics can be calculated reasonably accurately. Furthermore, it has been shown that the model is seed independent and its sensitive parameters identified (time, population size, average chronic patient visits and network failure probability). However, it was also shown to perform poorly at scale (large values of the critical parameters), was the data freshness metric performs erratically and could not be validated.

Given the tests performed, the boundary values in which the model performs adequately, is for populations of around 200, 000 or less, with a simulation study of 2 simulated years or less and approximately 40 facilities.

Question 3 - What are the implications of the answers to the preceding questions on the use of the programmed model as an evaluation tool?

The model can be used to compare how the two network algorithms would work in a given context, in terms of bandwidth and data freshness. The given contexts would have to include facilities that were similar to the Primary Healthcare Clinic model, servicing clients sequentially. The other limitation is that the simulation comparison would only be over a few years or less and, if dealing with larger patient volumes, sufficient hardware would be required. Besides these restrictions, the model will still allow much customisation and tailoring to the context.

Ultimately, the model could perhaps best be used to establish initial feasibility for a proposed network, which would establish at a high level if a given scenario will result in unrealistic levels of bandwidth usage or data freshness. Then, should the scenario appear feasible, further more detailed research could be done on specific questions with respect to the proposed solution.

In considering the risk of using this model for proposed network evaluation, Blaci's 9th Principle of Simulation VVT (Balci, Simulation Models, 1995) defines three errors for simulation studies which must be prevented:

- Type I error is when the results of a simulation study are rejected when they are actually sufficiently credible.
- Type II error is when incorrect simulation results are accepted as credible.

- Type III error is often performed when the study problem has been incorrectly formed and the simulation solves the wrong problem.

Considering these errors in the use of the NPR model for simulation studies, the only potential risk is that of type II, were the model has been applied to something beyond its scope (such as including full hospitals or running for 10 years or more) and then the results are accepted.

4.7 Summary

This chapter illustrated how the validation, verification and testing methodology was used to establish the scope and limitations of the NPR model. Several validation tests found that the model could replicate the behaviour of the EPR maintenance algorithms, the attending populations and the erratic network connections to facilities. Furthermore, it was demonstrated that the programmed model could reasonably accurately calculate bandwidth usage and data freshness for these networks.

It was shown that the model was seed independent and that its key parameters were those that affected the number of patient visits – namely population size, visit frequency, time – and those that represented the reliability of connections.

While performing adequately for small and medium networks, both the outputs and execution time performance of the model became unreliable when given long simulation times (10 years or more).

As such, the risk of using the model was reduced to the condition that the problem domain to be simulated must be inside the model's verified scope of time and network type.

Now that the accuracy and boundaries of the designed NPR model have been defined, Chapter 5 will demonstrate its use in a case study.

Chapter

5 Case Study

This chapter demonstrates how the model is used to evaluate two proposed solutions for a regional EPR network. In order to illustrate the application and benefits of this approach, a case study on the scalability of two NPR solutions will be demonstrated and reviewed.

5.1 Motivation and Desired Outcomes

Health districts in South Africa differ greatly in terms of both demand and number of facilities available. To consider the appropriateness of an EPR network, scalability is evaluated in relation to the Quality of Service (QoS) in networks of different sizes. The scale of a health network is represented by the number of transactions it must process – and often is a function of patient visits. The biggest determining factors for these transactions are the number of facilities and the population size. Therefore, this study will analyse the effect different scales of visit loads have on the Quality of Service (QoS) metrics the simulations produce for each network model, specifically considering the implications for cost and quality of care, in order to answer the following questions:

- How does the size of the network affect total bandwidth?
- How does network size influence data freshness?
- How does the volume of visits change the values of average and maximum bandwidth demand?

5.2 Methodology

The case study consists of several simulations that represent various scenarios from the real world. Each simulation produced quantitative metrics to assess the model's performance. The *centralised* and *de-centralised* models created earlier (now referred to as the UK and US models respectively) were run in each real world scenario, and their performance recorded.

The case study followed the process below:

1. *Define the study goal*: Identifying the clear objective of the study and the metrics or operation data required for analysis.
2. *Create study design*: Identifying the distinct simulations (one or more) that, given the operational data, provide the identified metrics.
3. *Perform simulations*: Saving the generated scalar and vector files as well as log output for metadata (execution times etc).
4. *Analyse results*: The output metrics of the study's simulations are compared and insights recorded.

5.2.1 Data Sets

The study required simple groupings of populations and facilities that represented various scales of district health regions in South Africa. Focusing on the input variables – population size (P) and number of facilities (F) – and their impact on visit volumes, four scenarios can be defined:

1. High concentrated visits – this is often typical of rural districts in South Africa, where a few PHC facilities can service a broad geographical region and its large population (i.e. low F, high P).
2. Low distributed visits – often typical of specialised point-of-service facilities (such as pharmacies) that only cover smaller part of the population (i.e. high F, low P).
3. Low concentrated visits – representing sparsely populated areas with the minimum required regional facilities (i.e. low F, low P).
4. High distributed visits – representing highly populated regions where patients have a broader range of facilities to visit (i.e. high F, high P).

While this is obviously a crude and discrete classification, it used to investigate the relationships between the network architectures and these two input parameters.

5.2.2 Simulations

The combination of small and large values for the *# Facilities* and *Population Size* input parameters were used to represent the four scenarios above. Running each of these against the two different network models yields the list of simulations inTable 19:

Table 19: Simulations for Scalability Study

#	NW	# Facilities	Pop size	Description	Example
5	US	3	100k	High concentrated visits	Rural KZN PHCs
6	UK	3	100k	High concentrated visits	Rural KZN PHCs
7	US	20	10k	Low distributed visits	Boland pharmacies
8	UK	20	10k	Low distributed visits	Boland pharmacies
9	US	3	10k	Low concentrated visits	Northern Cape pharmacies
10	UK	3	10k	Low concentrated visits	Northern Cape pharmacies
11	US	20	100k	High distributed visits	Gauteng PHCs
12	UK	20	100k	High distributed visits	Gauteng PHCs

The period for each simulation is one year.

5.2.3 Gathering results

From each simulation the main output metrics of cost (patterns in bandwidth demand and total bandwidth) and those of quality (data freshness) of each facility, population or server were gathered and aggregated by node type. The outcomes are displayed in the next section.

5.3 Results

5.3.1 Bandwidth Demand at Facilities

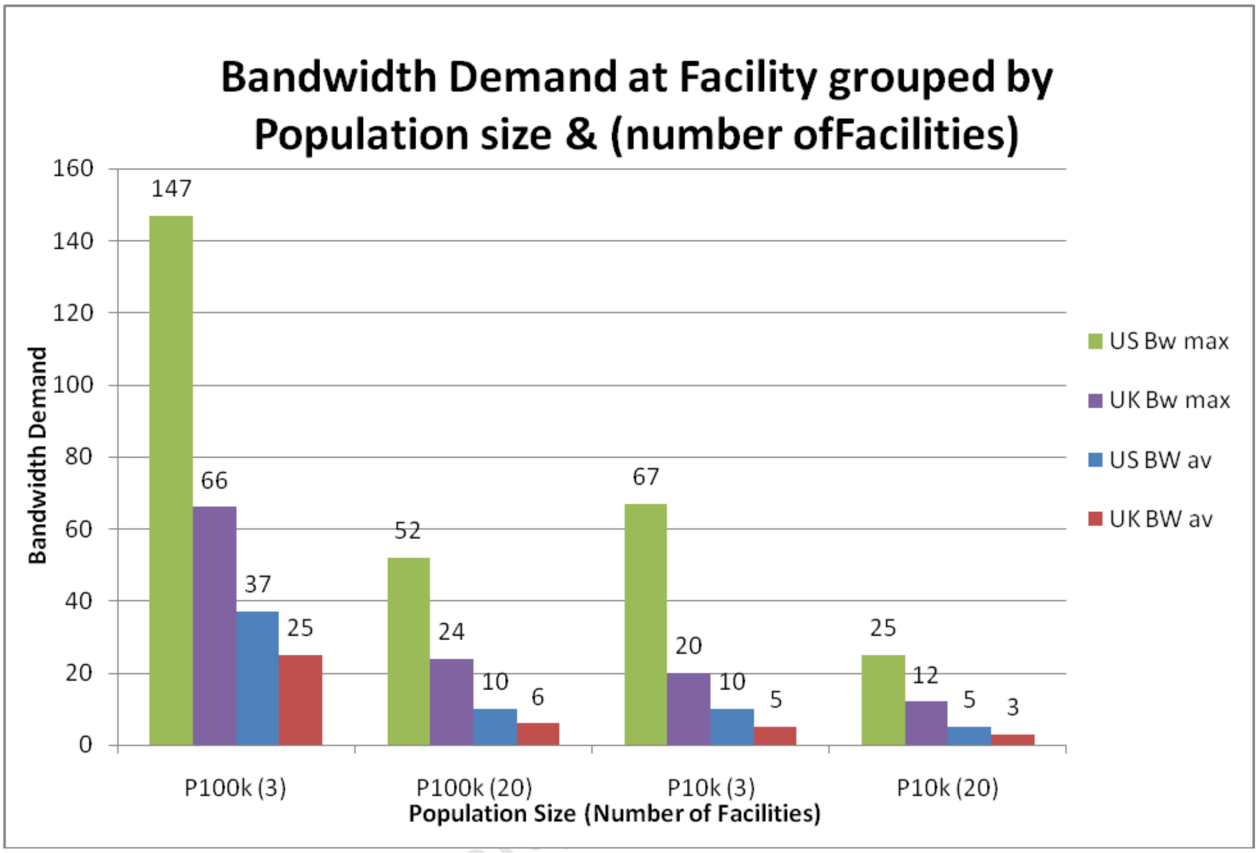


Figure 23: Bandwidth Demand at Facilities

Figure 23displays the results of the bandwidth demand at facilities, for the different simulations. Each grouping represents a real world simulation, the first group representing the scenario of a population of 100,000 visiting 3 facilities and so on. The histogram bars represent maximum and average bandwidth at an average facility for the two networks. Here it becomes possible to see trends across the simulations or between the two network models, and observations can be made.

The scenarios have represented two dimensions of growth or "scaling" – growth in population served and growth in number of Facilities in the network. Focusing on population growth, for 20 facilities, growing the population by 900% (from 10k to 100k) merely led to a 100% increase in the average & maximum bandwidth for both the US and UKnetworks. In contrast, for 3 facilities growing the population by 900% (from 10k to 100k) led to a 330% increase in maximum bandwidth for the UK model, and a 220% increase for the US. The increase in average bandwidth was also considerable with the UK increasing by 500% and the US by 370%. From this we also see that, for both models, an increase in population leads to a larger increase in average bandwidth more than maximum bandwidth. Furthermore, this impact is worse for fewer facilities – particularly in the UK model. Interestingly, this could indicate that while the US certainly has the most worrying worst case bandwidth demands (inferred from the maximum bandwidth results), it actually appears to perform better than the UK might at larger scales of network.

In thinking of costs, maximum and average bandwidth is more for the US network than the UK network in all scenarios for all facilities. This could indicate that US P2P networks would require more costly infrastructure to ensure this required bandwidth was available, or face reduced QoS on RTT given line capacity restrictions. However, the US Bandwidth appears to cope better with more facilities and UK with fewer. The cost implication could be that the US average performance scales well but suffers from outliers and could indicate the need for load sharing (when few facilities compete). The UK model handles increases in facilities well, but performance is always proportional to patient-facility ratios as UK suffers when this number is high. Ultimately, this result (focusing on the facility) could indicate that while US seems to scale more cheaply at the facility level, its QoS could be subject to severe fluctuations in worst case scenarios.

5.3.2 Bandwidth Demand at Server Nodes

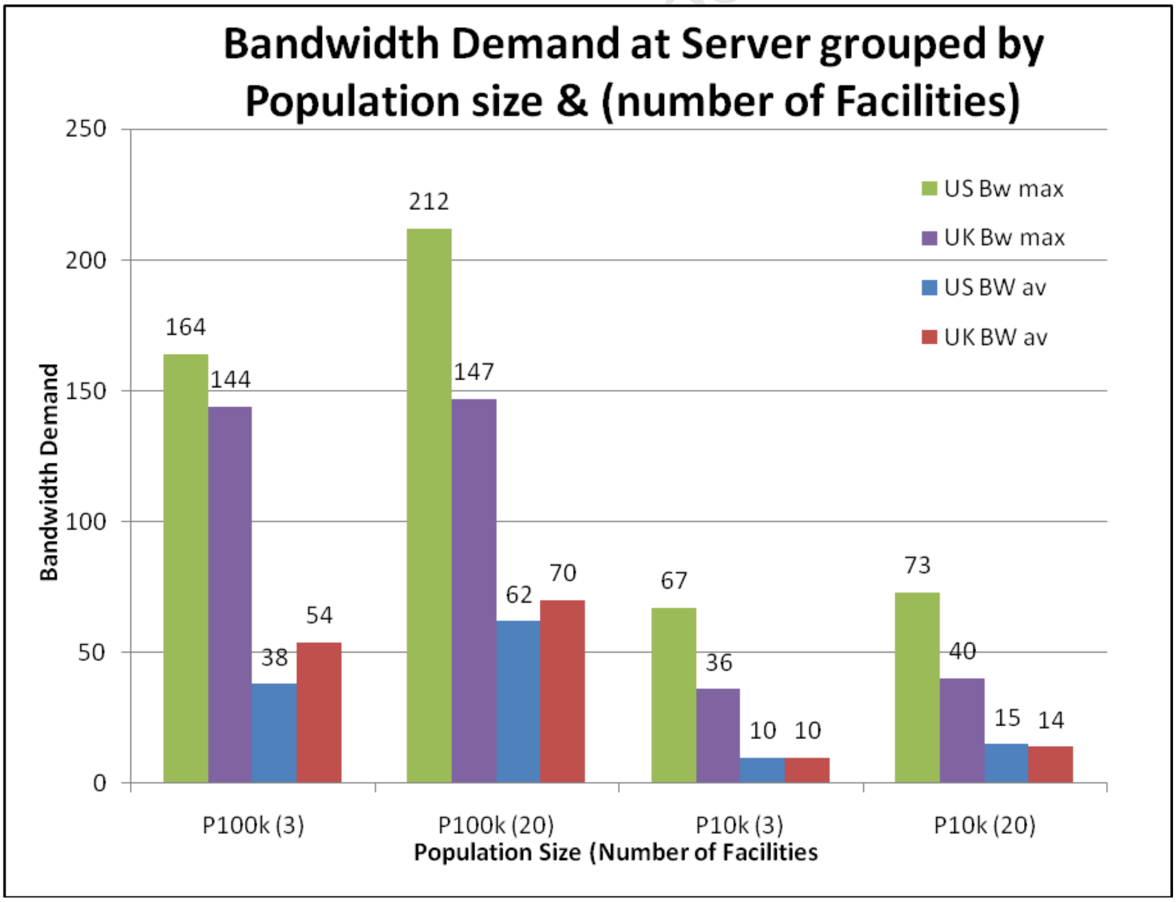


Figure 24: Bandwidth Demand at Servers

Figure 24displays the average and maximum bandwidth results – *at the server* – for each of the scenarios modelled. As opposed to the bandwidth demand at facilities (where the bandwidth value was some complex function of (P, N)), here bandwidth demand at the server is mostly proportional to the population and less sensitive to the number of facilities. This phenomenon is apparent in both network models where the coverage of the same population (namely 100k) yielded similar values in average Bandwidth across 3 or 20 facilities.

Notably it looks as though the increase in scale of population is more dramatically felt by UK networks. In increasing the population from 10k to 100k (900%) for three facilities, the average server bandwidth increased from 10 – 54 (440%), while the US average server bandwidth increased 300% between these two scenarios. Furthermore, a critical observation is that the average bandwidth demand at the server in the US network is *less or equal* in all scenarios against the UK network. This is in contrast to our observations above where the average bandwidth at facilities was *more* in all scenarios for the US network. The phenomenon of less demand on the US Peer-to-Peer server does match with conventional wisdom as the server is only polled on exceptional visits – not every visit like the UK centralised system.

5.3.3 Total Bandwidth usage at Facilities and Servers

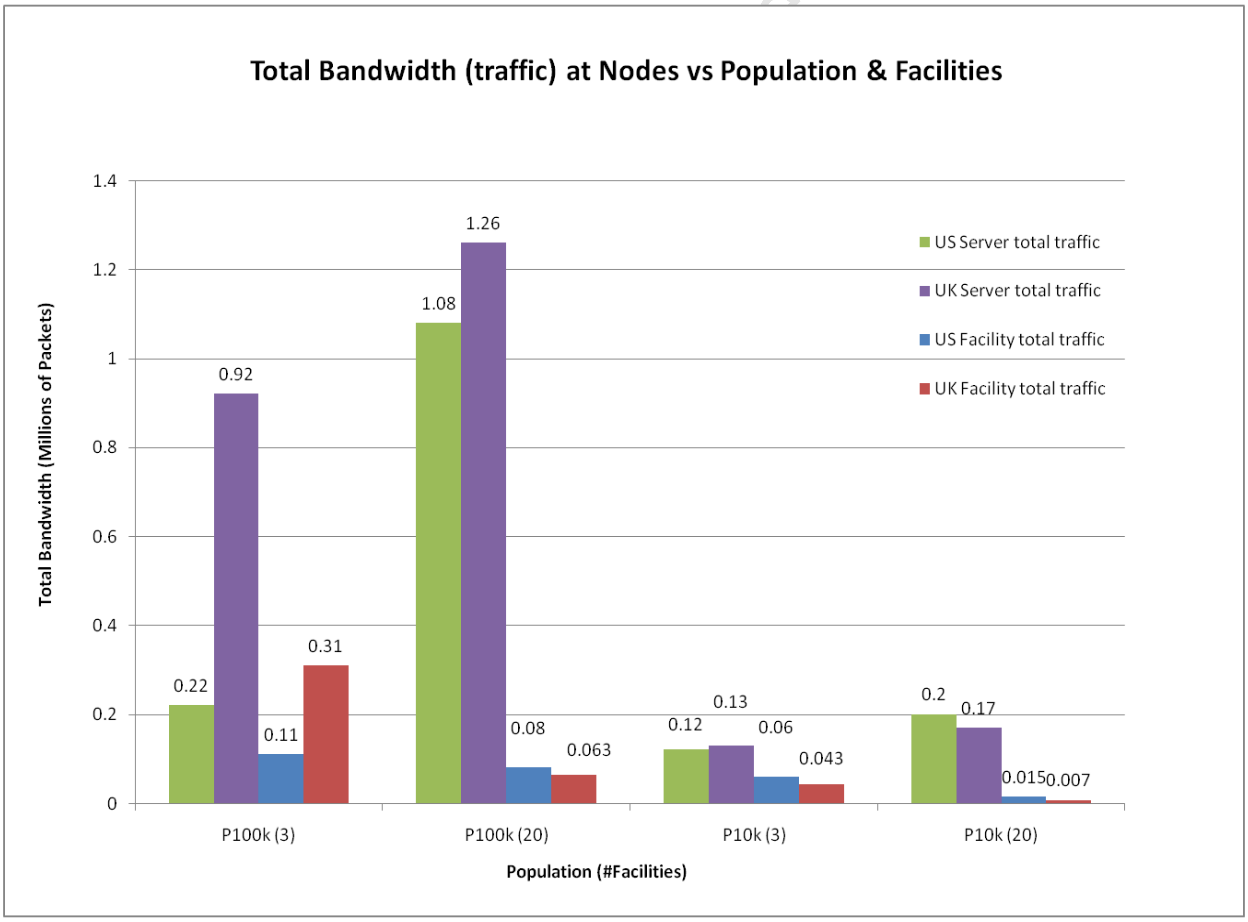


Figure 25: Total Traffic at All Node Types

Figure 25displays the total bandwidth used by a facility or server in each simulation scenario. Again, the UK network has by far the most bandwidth demand on the server in all scenarios. However, for a population of 100k, increasing the number of facilities from 3 to 20 yielded a 390% (from 0.22m to 1.08m) growth in total bandwidth traffic for the US network's server. The same increase in facilities yielded only a 36% increase in bandwidth traffic for the UK. In fact, given these initial observations, the effect of increasing the number of facilities on the server's bandwidth is actually inconclusive as a number of additional data points would be needed to extrapolate a trend.

Bandwidth demand for facilities in the US network appears be fairly consistent, whereas facilities in the UK network have again shown sensitivity to high patient-facility ratios – as displayed by the increase in both server and facility traffic when 3 facilities have to cover 100k population as opposed to 10k. Understandably the population size has a bigger effect on the total bandwidth usage than the number of facilities.

5.3.4 Data Freshness at Facilities

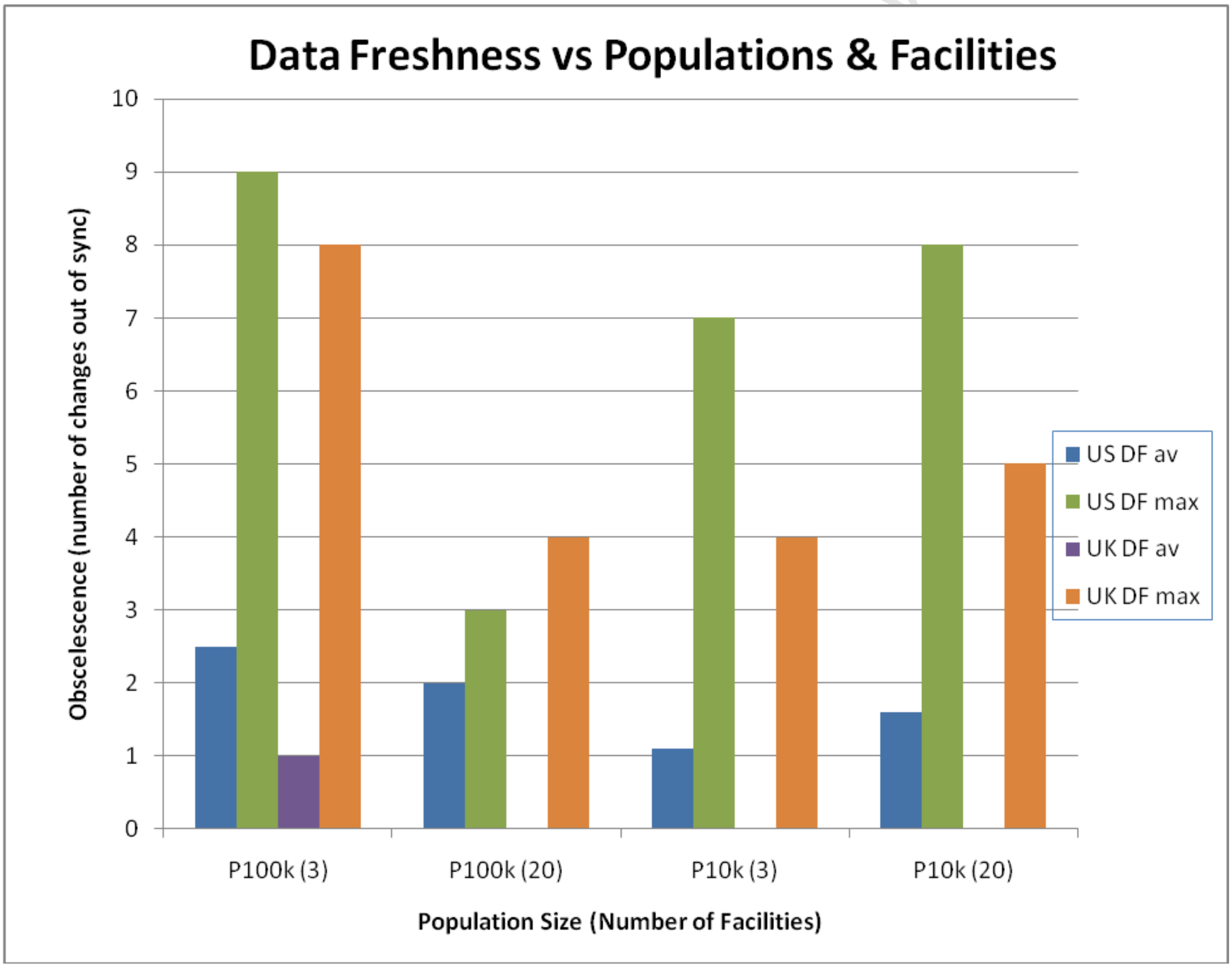


Figure 26: Data Freshness at Facilities

Figure 26 illustrates the results of the data freshness metric *obsolescence* at the facilities for each of the simulation scenarios. It should be noted that the absence of the "UK DF av" (purple) value in the second to fourth scenarios is actually a reflection that *this value is zero* – not missing. This then indicates that UK networks have an almost perfect average obsolescence (0 for 3 of the four scenarios), while the US averages around one or more obsolescence in the four scenarios. This piece of information confirms the fundamental difference between the two networks – i.e. that the UK provides the same record while US has several local copies that are not always coordinated.

Added to this, the US suffers from bad outliers – having the worst maximum obsolescencein all scenarios, though generally they seem to be random and do not follow an obvious trend that could be associated with network scales.

5.4 Analysis: Cost vs. Quality of Service

In the analysis of the four metrics observed in this case study (periodic facility bandwidth, periodic server bandwidth, total bandwidth usage at all nodes and data freshness), two predominant themes emerge:

1. The UK network model does not appear to scale as efficiently in terms of bandwidth as the US network. As networks grow larger (in terms of both population and number of facilities served), a UK network model would appear to become more costly to support.
2. The US network model, while appearing to scale well in terms of bandwidth costs, suffers in all metrics from outliers that could seriously affect the quality of service of supported applications. Added to this, its poor outputs of record accessibility and consistency (its scores in data freshness), lead to concerns over its reliability and usefulness in implementations and worse case scenarios.

Thus, given these two facts, we see that in terms of scalability there is a trade off between cost and Quality of Service between the two network architectures. The long term implications of this would need to be considered before any implementation project.

5.5 Scope and Limitations

This chapter aimed to demonstrate the feasibility of the modelling and simulation technique for evaluating healthcare networks. It has been demonstrated that the networks and real-world contexts can be modelled and their interactions simulated over a controlled period. These simulations in turn have yielded quantifiable metrics that can be used to infer implementation challenges and evaluations against other models.

The case study serves as an example of this process. Results were limited by the chosen depth of the original model and the limited number of scenarios pursued for the sake of demonstrative output. Along with more extensive simulation scenarios, parameter sensitivity and random seed generation could be taken into account when designing the study as a whole. These efforts, along with others introduced later in the future work section of the last chapter, would surely improve the quality of results for such experiments.

5.6 Summary

This chapter has demonstrated how the model can be used to evaluate network architectures in various scenarios, and given preliminary insight into the performance of the two network models under different conditions.

The two network models, the UK and US networks, were simulated over different network sizes to investigate their performance at different network sizes. The results showed that whilst the US model had significantly worse results in data freshness, its bandwidth costs scaled more gradually than that those of the UK network model.

The next chapter will conclude by synthesising the results and outcomes of the research, as well as proposing other case studies that the model could be used to evaluate.

Chapter

6 Conclusion

There are many potential solutions for implementing a National Patient Record in South Africa, yet little research has been completed on how these may perform under the constraints of a developing nation. The traditional approaches of evaluating networks - building pilot systems or using analytical projections - are predominantly expensive and difficult to validate. This research has proposed a third option – using modelling and simulation to understand the performance of NPR solutions in a given context.

6.1 Contribution

The study has shown that the NPR model developed is a feasible method for evaluating different NPR solutions in a given context. The model represents the key components of population, facilities, EPR data structures, network servers and contextual challenges such as downtime. These key components can be reused to model different network implementations. The model was used to represent two vastly different contemporary NPR solutions – namely the patient sharing mechanisms found in the UK's centralised NPR network and the United States' proposed peer-to-peer NPR network. Furthermore, the NPR model was shown to reasonably represent the local South African population and their patterns of Primary Health Care usage. This was achieved by creating a population of individually modelled patients, each of whose health care usage was determined by their patient condition – a once-off need, acute or chronic condition. Evaluating the cost and quality of the network's performance was also demonstrated by the inclusion of bandwidth usage and data freshness metrics, which could then be used for comparative analysis.

The model was used in a Case Study that investigated the performance of the UK and US models over different scales of local network size. The outcomes of the study showed that there was an inherent trade-off in bandwidth costs and data freshness quality between the two networks. The US model's bandwidth costs were observed to grow at a lower rate than the UK model's, as the population served increased. Alternatively, the UK model's data freshness values were found to be more consistent and suffer from fewer outliers than the US model for all networks. These results indicated that the UK model could be too costly at scale, and that the US model's difficulty to maintain the accuracy of the patient record may not be acceptable. Thus, the case study highlighted two areas of concern that the two NPR designs would need to improve on in the given contexts.

The development of this model, along with its evaluation and case study, has collectively shown that:

1. Different NPR models can be represented using the conceptual and programmed model developed. This can be used in simulation studies to compare NPR solutions in a given context.
2. Contextual challenges can be represented by line probability failure as well as node down time. Populations can be accurately modelled using patient type profiles, but the random distribution used to represent their arrival at the clinic does not yet have an impact on the performance of the network in the long run.
3. Total bandwidth and bandwidth demand can be used to assess the relative cost of implementing a particular network in a given context. Quality of service of the network algorithm can be represented by using a data freshness metric, such as obsolescence, to indicate how well the network is maintaining the accuracy of the Patient Records.

6.2 Scope and Limitations

The programmed model was shown to have a valid representation of the real world behaviour of the patient record algorithms as well as the visiting patterns of the patient population. The parameters of the population size, visit frequency and node failure proved to have the biggest influence over the outputs of the evaluation metrics. Alternatively, the careful modelling of patient arrival distributions did not produce any significant impact on the performance of the model.

While the model performed well for the small and medium networks tested, the integrity of the model was shown to break down under extreme values of population size or simulation time. Extremes in these inputs caused increased execution time and data freshness metric values that were difficult to justify.

Thus the confidence range of this specific implementation of the model is for simulation studies that are in the order of one year and populations in the order of one million. Furthermore, the model only represents groups of primary healthcare facilities servicing a fixed population, while in the real world there are more kinds of facilities and populations are seldom fixed. Ultimately, the current implementation of the model could not be used for the analysis of a full national multi-year simulation. However, it can be used to analyse specific conditions within a specific period – such as a disease outbreak, or resource failures at primary healthcare facilities in a particular region.

As suggested in Chapter 4, the limitation of the performance of this specific simulation was likely limited by programming errors or insufficient allocation of hardware resources for the experiments. A paper by (Mallanda & Suri, 2005) showed that in fact OMNet++ out performed NS2 in terms of execution time and memory consumption using networks of over 2000 nodes, indicating that OMNet++ should be able to scale more gracefully for most implementations.

6.3 Future Work

This work has provided a platform for further research into evaluating network solutions in the developing context. The analysis of the model itself as well as the case study produced many areas for improvement and extension, and also brought up interesting new questions that the technique could potentially address.

Future research can be done in three main areas: firstly, work that uses the existing model in further case studies to compare network performance in certain contexts; secondly, extensions to the model to allow for new problems to be addressed; and lastly, improvements to the programmed tool itself to increase its accessibility and use in additional research studies.

The current NPR model can be used in many case studies to evaluate the US and UK network architectures' performance in specific scenarios – as demonstrated in the case study that focused on their comparative performance in different sizes of network. Other contextual scenarios could be represented by the model to provide new ways to evaluate the two network models' performance. Scenarios such as a local as disease outbreak in the population, or a more in-depth study on how the networks perform with solely rural facilities, could be produced by providing new parameter configurations to the current model. For best results, these scenarios would need to be guided by real world problems or questions around potential network performance.

The second area for future work consists of the many possible extensions to the model. Depending on the research focus, certain areas of the model could be extended using new data sets to refine the model's representation of real-world behaviour. Some potential extensions include:

- Extend the facility model to represent more complex facility types – perhaps those that have multiple instances of the *point-of-service*sub-module as well as multiple input queues, such as large hospitals or clinics. Furthermore, research on real world data could provide blueprints for the processing models and capacity of facility types such as provincial hospitals, emergency wards, pharmacies and laboratories. Such additions would provide a more accurate picture of the type of demand placed upon an NPR network providing EPR synchronisation between all facilities.

- Extensive research could also be done into improving the population model to better replicate the real-world behaviour of individual patients. The patient types presented here could each be researched using data from real-world systems on the particular behaviour patterns of patients. Such data could potentially extend the patient type profile to include statistics on facility affinity of patients well as trends in birth and deaths.

- More NPR algorithms can be modelled and added to the programmatic model. Particularly algorithms and designs that look to address some of the problems noted with the UK and US models could be developed and then simulated in the same contexts. An example addition taken from the algorithms used in certain Point-of-Sale (POS) credit card networks could be: “can the store and forward routines used by POS networks maintain the UK model’s data quality whilst improving cost and scalability?”

The third area of future work could address the current high complexity of performing case studies using the NPR model. New improvements in the latest OMNeT++ simulation tool can be used to create a simpler workflow for designing and performing the simulation study. Furthermore, the programmed model itself could be refactored to provide a simpler interface to the study designer for configuration.

Ultimately this study has demonstrated the base feasibility of the simulation technique in evaluating the different network solutions. Further research should tailor and extend the model based on the influence of real-world problems and challenges in this domain.

6.4 Reflection

This research closely followed Law’s (Law, Designing a simulation study, 2003) methodology for simulation studies and some key insights from this process include:

1. Performing evaluator simulation studies of this kind requires a multi-disciplinary approach. Skills and research into technical areas such as programming, modelling and simulation need to be supplemented with skills and resources from public health research or the domain in question.

2. The iterative approach proved crucial in producing a model that was closer to the expected real world behaviour (as illustrated by the population).

3. Unit testing was found to be an innovative and helpful technique in maintaining the validity of the programmed model as the conceptual model evolved.

4. Complexity of implementation is fairly high and this could be a barrier to accessing the technique. However, advancements in the usability of the OMNeT++ tool along with reusable protocol frameworks could aid further research into its general usability.

6.5 Final Thoughts

The challenge of building NPRs continues to be a complex and difficult problem to implement and design, as illustrated by the recent scaling back of the £12bn NHS IT system (BBC News, 2009). This research has shown that modelling potential NPR designs and their associated contexts is a feasible method of evaluation. Further research can build on this to work and potentially elevate it to be used in collaboration with other techniques when considering national (Molkdar, Burley, & Wallington, 2002) implementations.

Bibliography

(Abiteboul, 2004) Abiteboul, S., Alexe, B., Benjelloun, O., Cautis, B., Fundulaki, I., Milo, T., and Sahuguet, A. 2004. An electronic patient record "on steroids": distributed, peer-to-peer, secure and privacy-conscious. In *Proceedings of the Thirtieth international Conference on Very Large Data Bases - Volume 30* (Toronto, Canada, August 31 - September 03, 2004). M. A. Nascimento, M. T. Özsu, D. Kossmann, R. J. Miller, J. A. Blakeley, and K. B. Schiefer, Eds. Very Large Data Bases. VLDB Endowment, 1273-1276.

(Accenture, 2006) Accenture. 2006. *Achieving High Performance in Health Care – Insights into the Introduction of Electronic Health Records in South Africa.*Retrievied 2007, Sep from Website of Accenture: http://www.accenture.com/NR/rdonlyres/15D767DA-7F3F-4AE6-ABDD-8260DA7BFFC5/0/eHRbrochureJune06.pdf

(Arnold, 2002) Arnold, Ken. 2002. *Designing Distributed Systems,* Interview with Artimer Developer part III. http://www.artima.com/intv/distrib.html

(Au-Yeung, 2006) Au-Yeung, S.W.M., Harrison, P.G., & Knottenbelt, W.J. 2006. A queuing network model of patient flow in an accident and emergency department, Department of Computing, Imperial College of London, August 15, 2006

(Balci, VVT Life Cycle, 1994) Balci, O. 1994. Validation, verification, and testing techniques throughout the life cycle of a simulation study. In *Proceedings of the 26th Conference on Winter Simulation* (Orlando, Florida, United States, December 11 - 14, 1994). M. S. Manivannan and J. D. Tew, Eds. Winter Simulation Conference. Society for Computer Simulation International, San Diego, CA, 215-220.

(Balci, Simulation Models, 1995) Balci, O. 1995. Principles and techniques of simulation validation, verification, and testing. In *Proceedings of the 27th Conference on Winter Simulation* (Arlington, Virginia, United States, December 03 - 06, 1995). C. Alexopoulos and K. Kang, Eds. Winter Simulation Conference. IEEE Computer Society, Washington, DC, 147-154

(Balsamo, 1998) Balsamo, S., Inverardi, P., and Mangano, C. 1998. An approach to performance evaluation of software architectures. In *Proceedings of the 1st international Workshop on Software and Performance* (Santa Fe, New Mexico, United States, October 12 - 16, 1998). WOSP '98. ACM Press, New York, NY, 178-190. DOI= http://doi.acm.org/10.1145/287318.287354

(BBC News, 2009) BBC News. (2009, December). *Troubled 12bn NHS IT System to be Scaled Back.*From BBC News Website: http://news.bbc.co.uk/2/hi/uk_news/politics/8397854.stm

(Beck, 2004) Beck, K. and Andres, C. 2004. *Extreme Programming Explained: Embrace Change (2nd Edition).* Addison-Wesley Professional.

(Berg, 2004) Berg, M. 2004. *Health Information Management: Integrating information technology in health care work*. Published by Routledge 2004. ISBN 0-145-31519-0

(Beyer, 2004) Beyer, M., Kuhn, K. A., Meiler, C., Jablonski, S., and Lenz, R. 2004. Towards a flexible, process-oriented IT architecture for an integrated healthcare network. In *Proceedings of the 2004 ACM Symposium on Applied Computing* (Nicosia, Cyprus, March 14 - 17, 2004). SAC '04. ACM Press, New York, NY, 264-271. DOI= http://doi.acm.org/10.1145/967900.967958

(Bouzeghoub, 2004)Bouzeghoub, M. 2004. A framework for analysis of data freshness. In *Proceedings of the 2004 international Workshop on information Quality in information Systems* (Paris, France). IQIS '04. ACM, New York, NY, 59-67.

(Braa & Hedberg, 2000) Braa J., Hedberg C. 2000. Developing District-based Health Care Information Systems: The South African Experience. http://iris23.htu.se/proceedings/PDF/59final.PDF, accessed May 2006

(Canner, 1992) Canner, J. K., Chiang, Y., and Javitt, J. C. 1992. A Monte Carlo based simulation network model for a chronic progressive disease: the case of diabetic retinopathy. In *Proceedings of the 24th Conference on Winter Simulation* (Arlington, Virginia, United States, December 13 - 16, 1992). WSC '92. ACM, New York, NY, 1041-1049.

(CIA, 2007)CIA.2007. *CIA World Factbook*. Retrieved 2007 fromCIA Website https://www.cia.gov/cia/publications/factbook/print/us.html.

(Coyne, 2008) Coyne, M. E., Graham, S. R., Hopkinson, K. M., and Kurkowski, S. H. 2008. A methodology for unit testing actors in proprietary discrete event based simulations. In *Proceedings of the 40th Conference on Winter Simulation* (Miami, Florida, December 07 - 10, 2008). S. Mason, R. Hill, L. Mönch, and O. Rose, Eds. Winter Simulation Conference, 1012-1019.

(Dudenhoffer, 2002)Dudenhoeffer, D. D., Permann, M. R., and Sussman, E. M. 2002. General methodology 3: a parallel simulation framework for infrastructure modeling and analysis. In *Proceedings of the 34th Conference on Winter Simulation: Exploring New Frontiers* (San Diego, California, December 08 - 11, 2002). Winter Simulation Conference, 1971-1977.

(Fall, & Varadhan, 2004). Fall, K., Varadhan, K., (2004). NS-2 Network Simulator. Berkely: Universtiy of California.

(Franke, 2009)Franke, U., Flores, W. R., and Johnson, P. 2009. Enterprise architecture dependency analysis using fault trees and Bayesian networks. In *Proceedings of the 2009 Spring Simulation Multiconference* (San Diego, California, March 22 - 27, 2009). Spring Simulation Multiconference. Society for Computer Simulation International, San Diego, CA, 1-8.

(Gartner, 2007) Gartner. (2007, May 31). *Summary of the NHIN Prototype*

Architecture Contracts: A Report for the Office of the National Coordinator for Health IT. Retrieved from the Website for the Dept. of Health & Human Services: www.hhs.gov—summary_report_on_nhin_Prototype_architectures.pdf.

(Health Systems Trust, 1998) Ntuli, Antoinette (ed). 1998. *The South African Health Review 1998.* Health Systems Trust. Available at URL: http://www.hst.org.za/publications/210/

(Health Systems Trust, 2005) Ijumba, Petrida and Barron, Peter [eds]. 2005. *South African Health Review 2005*. Health Systems Trust (August 2005). Available at URL: http://www.hst.org.za/publications/682

(Health Systems Trust, 2004) Reagon, G., Irlam J, Levin J. 2004. *The National Primary Health Care Facilities Survey 2003*. Health Systems Trust (June 2004). Available at URL: http://www.hst.org.za/publications/617/

(Heugh, 2007) Heugh, K. 2007. Language and literacy issues in South Africa. In *Global issues in language, education and development: perspectives from postcolonial countries.* Clevedon: Multilingual matters. pp. 187-286.

(International Institute of Business Analysis, 2008) International Institute of Business Analysis. 2008.*Business Analysis Body of Knowledge (Version 1.6)*.

(Jones, 2004) Jones, T. M. 2004. National Infrastructure for eHealth: Considerations for Decision Support. Studies in Health Technology and Informatics, Volume 100, 2004. *E-Health - Current Situation and Examples of Implemented and Beneficial E-Health Applications*. ISBN 978-1-58603-448-1

(King, 1983) King, J. L. 1983. Centralized versus decentralized computing: organizational considerations and management options. *ACM Comput. Surv.* 15, 4 (Dec. 1983), 319-349.

(Law, Designing a simulation study, 2003)Law, A. M. and McComas, M. G. 2001. Building valid models: how to build valid and credible simulation models. In*Proceedings of the 33nd Conference on Winter Simulation* (Arlington, Virginia, December 09 - 12, 2001). Winter Simulation Conference. IEEE Computer Society, Washington, DC, 22-29.

(Law, Building Valid Models, 2003) Law, A. M. 2003. Designing a simulation study: how to conduct a successful simulation study. In *Proceedings of the 35th Conference on Winter Simulation: Driving innovation* (New Orleans, Louisiana, December 07 - 10, 2003). Winter Simulation Conference. Winter Simulation Conference, 66-70.

(Mallanda & Suri, 2005) Mallanda, C., Suri, A., 2005. Simulating Wireless Sensor Networks with OMNeT++. Louisiana State University, LSU Simulator, Version 1.

(Medical Records Institute, 2002) Medical Records Institute. 2002. *Impact of Hospital Computer Systems on Resident Work Hours.* Retrieved 2007 from the Website of the Medical Records Institute: http://www.medrecinst.com/libarticle.asp?id=26

(Microsoft) Microsoft Developer's Network. *What is a Transaction?* [Online] [Cited: 6 February 2010.] http://msdn.microsoft.com/en-s/library/aa366402(VS.85).aspx

(Molkdar, Burley, & Wallington, 2002)Molkdar, D. Burley, S. Wallington, J. 2002.*Comparison Between Simulation and Analytical Methods of UMTS Air Interface Capacity Dimensioning*. Vehicular Technology Conference, 2002. Proceedings. VTC 2002-Fall. 2002 IEEE 56th

(National Committee on Vital and Health Statistics, 2006)National Committee on Vital and Health Statistics.(2006, September).*Minimum but Inclusive Functional Requirements Needed for the Initial Definition of a NHIN.*Retrieved 2007 from Health and Human Services Website: http://www.ncvhs.hhs.gov/060831p1.pdf

(National Energy Regulator of South Africa, 2008)National Energy Regulator of South Africa. 2008. *Inquiry into the National Electricity Supply Shortage and Load Shedding.* Retrieved from NERSA Website: http://www.nersa.org.za/UploadedFiles/RegulatorsDecisions/LSEnergyRegulatorReportByELS14May2008V3Rev8SSMFinalpdf.pdf

(National Health Service, 2006) National Health Service. (2010).*Standards Guidence – HL7*. Retrieved 2010 from Connecting for Health Website: http://www.connectingforhealth.nhs.uk/industry/step/guidence

(New Zealand Health Information Service) New Zealand Health Information Service (n.d.).*Guide to NZHIS National Collections.* From New Zealand Health Information Service Website: http://www.nzhis.govt.nz/moh.nsf/pagesns/60

(Northrop Grumman, 2006)Northrop Grumman. 2006.*Nationwide Health Information Network – The Northrop Grumman Architectural Approach.* Retrieved 2007 from Website for the Dept. of Health & Human Services: http://www.hhs.gov/healthit/documents/NGC_NHIN_Forum.pdf

(Ohio State University, 2001) Ohio State University Health System. 2001. The Design and Implementation of a Computerized Patient Record at the Ohio State University Health System – a Success Story. http://www.himss.org/content/files/davies_2001_osuhs.pdf

(Puigjaner, 2003) Puigjaner, R. 2003. Performance modelling of computer networks. In *Proceedings of the 2003 IFIP/ACM Latin America Conference on Towards A Latin American Agenda For Network Research* (La Paz, Bolivia, October 03 - 05, 2003). LANC '03. ACM, New York, NY, 106-123.

(Ringholm, 2007) Ringholm. 2007.*The Spine Architechture, an English National Programme (Whitepaper).* Retrieved2007 from: http://www.ringholm.de/docs/00970_en.htm. accessed June 2007.

(Sargent, Verification and Validation of simulation models, 1998)Sargent, R. G. 1998. Verification and validation of simulation models. In *Proceedings of the 30th Conference on Winter Simulation* (Washington, D.C., United States, December 13 - 16, 1998). D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, Eds.

Winter Simulation Conference. IEEE Computer Society Press, Los Alamitos, CA, 121-130.

(Sargent R. G., 2000) Sargent, R. G., Glasow, P. A., Kleijnen, J. P., Law, A. M., McGregor, I., and Youngblood, S. 2000. Strategic directions in VV&A; research: strategic directions in Verification, Validation, and Accreditation research. In *Proceedings of the 32nd Conference on Winter Simulation* (Orlando, Florida, December 10 - 13, 2000). Winter Simulation Conference. Society for Computer Simulation International, San Diego, CA, 909-916.

(Sargent, Model verification, 2007) Sargent, R. G. 2007. Verification and validation of simulation models. In *Proceedings of the 39th Conference on Winter Simulation: 40 Years! the Best Is Yet To Come* (Washington D.C., December 09 - 12, 2007). Winter Simulation Conference. IEEE Press, Piscataway, NJ, 124-137.

(South African Department of Health, 2001) South African Department of Health. 2001. *National Telemedicine Programme and Priorities in South Africa.* Retrieved 2006 from the Website of the Department of Health: http://www.doh.gov.za/programmes/tele/july01.html

(Swisher, 1997) Swisher, J. R., Jacobson, S. H., Jun, B., and Balci, O. 1997. Simulation of the Queston Physician Network. In *Proceedings of the 29th Conference on Winter Simulation* (Atlanta, Georgia, United States, December 07 - 10, 1997). S. Andradóttir, K. J. Healy, D. H. Withers, and B. L. Nelson, Eds. Winter Simulation Conference. IEEE Computer Society, Washington, DC, 1146-1154.

(U.S. Census Bureau, 2004) U.S Census Bureau. 2004. *Income, Poverty, and Health Insurance Coverage in the United States*.U.S Census Bureau.

(U.S. Census Bureau, 2005) U.S Census Bureau. 2005. *Facts for Features: National Nurses and Natinoal Hospital Week*.US Census Bureau.

(U.S. Dept. of Health & Human Services, 2005) U.S. Department of Health & Human Services services. (2005, Nov). *News Release: HHS Awards Contracts to Develop Nationwide Health Information Network.* Retrieved 2007 from Website of the Dept. of Health & Human Services: http://www.hhs.gov/news/press/2005pres/20051110.html

(Varga, 2001) Vargas. 2001. OMNET++ Discrete Event Simulation System. *European Simulation Multiconference*

(Waupotitsch, 2006) Waupotitsch, R., Eidenbenz, S., Smith, J. P., and Kroc, L. 2006. Multi-scale integrated information and telecommunications system (MIITS): first results from a large-scale end-to-end network simulator. In *Proceedings of the 38th Conference on Winter Simulation* (Monterey, California, December 03 - 06, 2006)

Appendix A

The Patient Class:

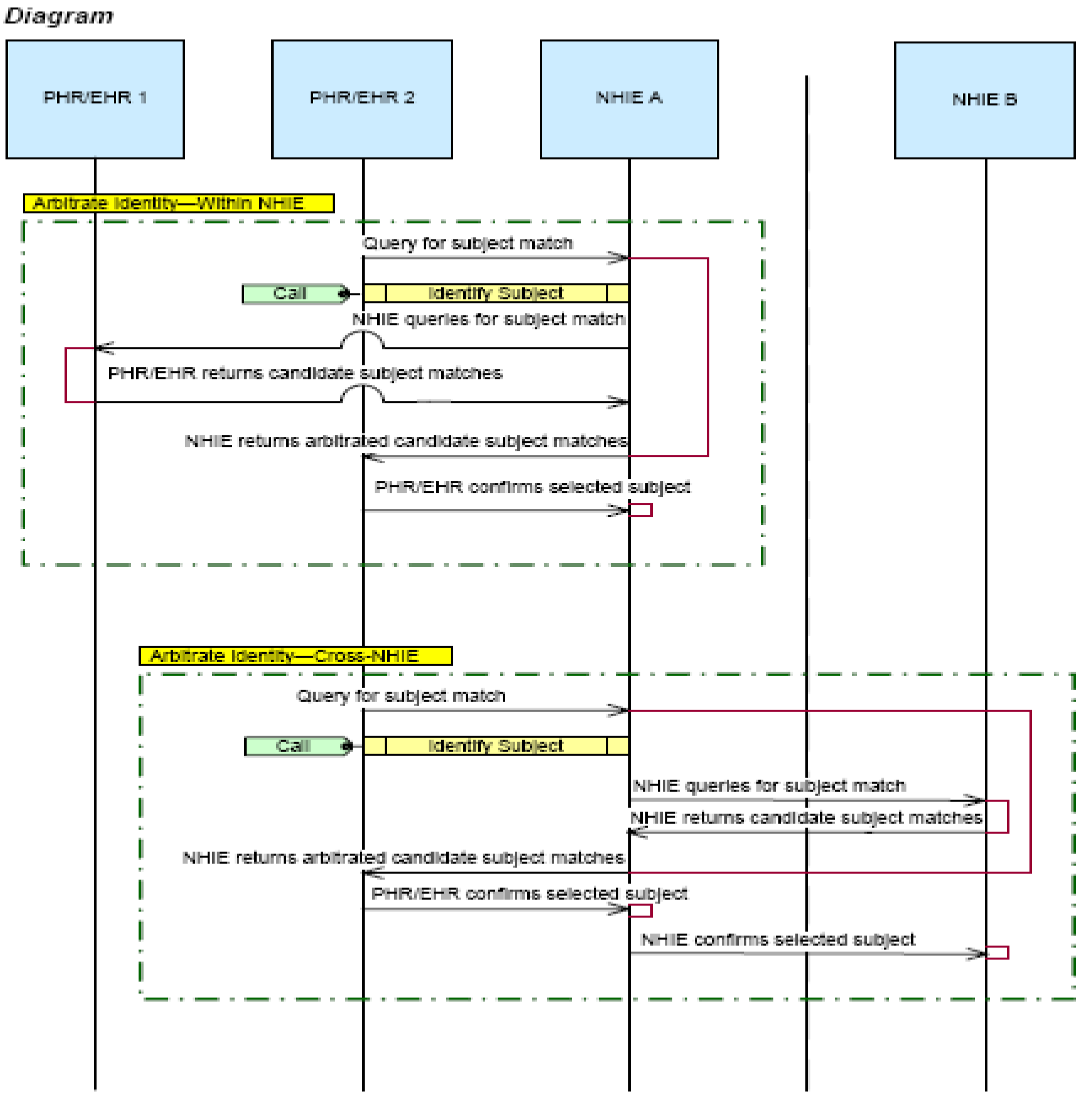
```
class Patient{

      public:
      conststaticint TYPE_NEVER = 0;
      conststaticint TYPE_SINGLE = 1;
      conststaticint TYPE_ACUTE = 2;
      conststaticint TYPE_CHRONIC = 3;
      conststaticint MEAN_VISITS_ACUTE = 5;
      conststaticint MEAN_VISITS_CHRONIC = 12;

       Patient(bool debug);
      void setPatientType(int i);
      int getNumberOfVisits();
      void generateVisits(long simDuration, int facilities);
      long checkInWindow(long delay, long simDuration, long 
dayWindow);
      long getVisitTime(int i);
      int getVisitFacility(int i);
      void setAcuteVisitsMean(long mean);
      void setAcuteVisitsStddev(long stddev);
      void setFacilityRandomness(double randomness);

      private:
      bool debug;
      int type;
      int facility;
       std::vector<long> visitTimes;
       std::vector<int> visitFacility;
      long acuteVisitsMean;
      long acuteVisitsStddev;
      double facilityRandomness;
};
```

The RecordRepository Class:

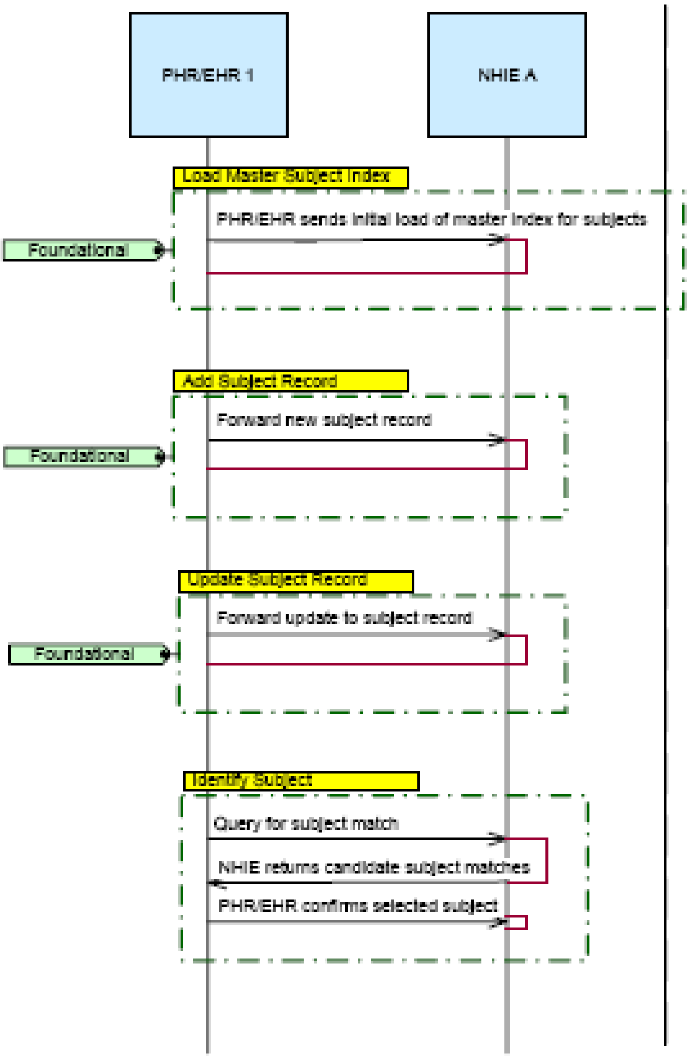
```
class MasterRecord
{

private:
    hash_map<string,int> PImap;

public:
    MasterRecord();
virtual ~MasterRecord();
void update_MasterRecord(string uid);
int getVisits(string uid);
    string toString();

staticconstbool debug = true;

`
```

Appendix B: NHIN sequence diagrams:

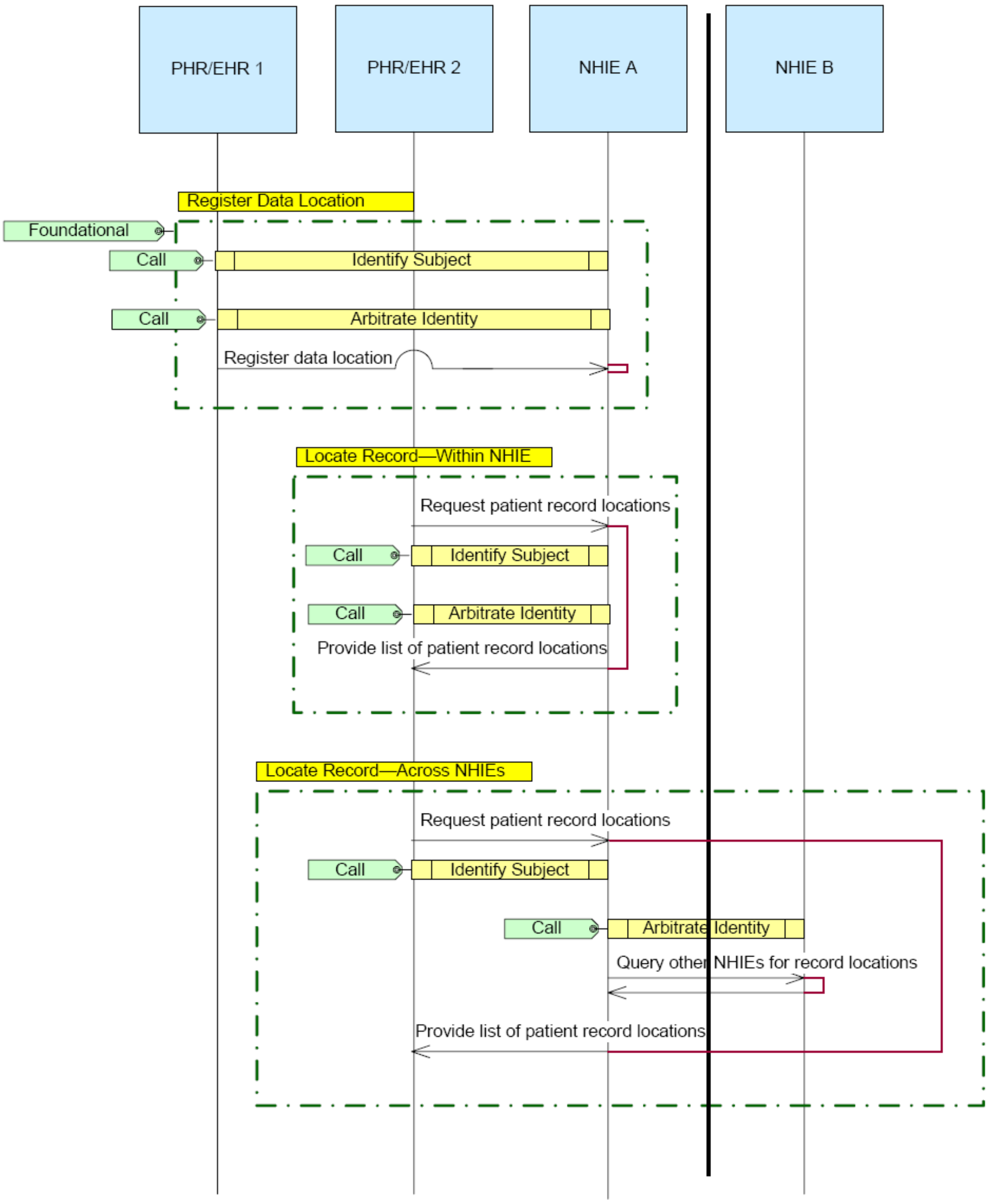
Arbitrate Identify of patient



Diagram


PHR/EHR 1
NHIE A
Load Master Subject Index
PHR/EHR sends initial load of master index for subjects
Foundational
Add Subject Record
Forward new subject record
Foundational
Update Subject Record
Forward update to subject record
Foundational
Identify Subject
Query for subject match
NHIE returns candidate subject matches
PHR/EHR confirms selected subject

Diagram


PHR/EHR 1
PHR/EHR 2
NHIE A
NHIE B
Register Data Location
Foundational
Call
Identify Subject
Call
Arbitrate Identity
Register data location
Locate Record—Within NHIE
Request patient record locations
Call
Identify Subject
Call
Arbitrate Identity
Provide list of patient record locations
Locate Record—Across NHIEs
Request patient record locations
Call
Identify Subject
Call
Arbitrate Identity
Query other NHIEs for record locations
Provide list of patient record locations