The Nuffield Trust
FOR RESEARCH AND POLICY
STUDIES IN HEALTH SERVICES

Predicting social care costs: a feasibility study

Martin Bardsley, John Billings,
Ludovic Jean Chassin, Jennifer Dixon,
Elizabeth Eastmure, Theo Georghiou,
Geraint Lewis, Rhema Vaithianathan,
Adam Steventon

PREDICTING SOCIAL CARE COSTS

A FEASIBILITY STUDY

Martin Bardsley

John Billings

Ludovic Jean Chassin

Jennifer Dixon

Elizabeth Eastmure

Theo Georghiou

Geraint Lewis

Rhema Vaithianathan

Adam Steventon

The Nuffield Trust
FOR RESEARCH AND POLICY
STUDIES IN HEALTH SERVICES

ABOUT THE NUFFIELD TRUST

The Nuffield Trust is a charitable trust carrying out research and health policy analysis on health services. Our focus is on the reform of health services to improve the efficiency, effectiveness, equity and responsiveness of care.

The Nuffield Trust
59 New Cavendish Street
London W1G 7LP

Telephone: 020 7631 8450
Fax: 020 7631 8451

Email: info@nuffieldtrust.org.uk
Website: www.nuffieldtrust.org.uk

Charity number 209201

This research, which was originally published in the journal *Age and Ageing*, is reported here by permission of the British Geriatrics Society:

Bardsley M, Billings J, Dixon J, Georghiou T, Lewis GH and Steventon A (2011) 'Predicting who will use intensive social care: case finding tools based on linked health and social care data', *Age and Ageing* 40(02).

CONTENTS

ACKNOWLEDGEMENTS

This project would not have been possible without the support and enthusiasm of the primary care trusts, local authorities and the care trust at five sites in England that provided us with the data we used for modelling. Each of the five sites (Bristol, Croydon, Devon, Torbay and Western Cheshire) also worked with us to ensure that our findings made sense to front-line professionals.

We are also grateful to the project reference group – and to Ray Beatty and Guy Robertson in particular – for their encouragement, insight and guidance.

Thanks also to:

- Health Dialog UK for the use of their pseudonymisation tool in this project and for adaptation of their business planning tool.
- Experian UK for allowing us to test their Mosaic™ variables, and to Emily Sparks for her helpful advice.
- Johns Hopkins University for allowing us to test the impact of the ACG™ system on our predictions, and advice from Steve Such.

EXECUTIVE SUMMARY

The costs of caring for people with complex social care and health care needs are set to rise in the UK over the coming years. As more people live with long-term medical conditions, it will become increasingly important to find ways to find ways of helping local councils and health services to take earlier action to support people to remain independent and stay in their own homes.

This report describes a study that explored whether statistical models can be used to predict an individual person's future need for intensive social care. Aside from the predictive models we developed, this work generated important lessons about the potential of linked health and social care data to support policy analysis and to guide the planning and commissioning of services.

KEY POINTS

- Although health and social care services interact in many ways for millions of people, their information systems tend to be discrete and distinct. This research has shown how it is possible to link routine data from health and social care information systems in a way that protects individuals' identities.
- Within health care, predictive modelling is increasingly used as a strategy to identify people at high risk of future unplanned hospital admission, and so target preventive care. Such approaches have not previously been tested with respect to social care. Predictive models have the potential to provide a better experience for service users and to offer more cost-effective care.
- This project has shown that it is possible to construct predictive models for social care. The next stage will be to see how these models might fit into everyday working practice.
- The predictive accuracy of our models is comparable to some of the models used by the NHS to predict hospital admissions. We suggest that it will be important to pilot and evaluate the use of these tools in practice, across a range of sites.
- Linked person-level information has the potential to improve the quality of care services – whether through improved identification of high-risk individuals, comparative performance measures, service evaluations or budget-setting. At a time when individual budgets and personalisation are seen as important, the need to collate and analyse information of this type seems ever more pressing.
- The quality of data about individual health care use has improved considerably over the past decade. Now a step change is needed to ensure that information about social care services improves in the same way. This will require strategies to improve the coding, collection and sharing of data in ways that protect confidential information.

1. INTRODUCTION

The social care costs and healthcare costs of people with complex needs are set to rise steeply in the UK over the medium term. This is due to the ageing population and the growing number of people living with long-term medical conditions. Both types of cost are highly skewed across the population, with a small number of individuals accounting for most of the expenditure. Being able to identify these people would be helpful so that they could be offered targeted, effective support and preventive care aimed at promoting independent living. Such 'upstream' investment has the potential to yield substantial net savings 'downstream' if the start of intensive social care could be delayed or avoided.

Over the last few years, many NHS organisations in England have started using predictive tools to work out which individuals in a given population are at risk of unplanned admission to hospital. Predictive models use historic patterns in the population's data to make predictions at the individual level.[1] The Department of Health commissioned two such tools for the NHS in England, the Patients at Risk of Re-hospitalisation (PARR) and the 'Combined Model', which are able to identify with reasonable accuracy those people in a population who are at risk of unplanned hospital admission or readmission in the forthcoming year. The tools use de-identified ('pseudonymous') administrative data to generate risk scores at the individual level, which are made available to GP practices. GPs are then able to offer interventions such as 'case management' (community matrons etc.) and 'disease management' (health coaching etc.) to high-risk patients, aimed at mitigating the risk of future hospitalisation.

This study explores whether similar predictive tools are feasible for social care. Emergency hospital admission and admission to a care home are analogous, in that both events are typically:

- unwelcome to the person concerned
- costly to society
- recorded in routine electronic data
- sometimes preventable.

There is high-quality evidence from the literature that certain interventions, such as domiciliary multi-dimensional geriatric assessment, can successfully prevent or delay care home admissions.[2] However, such programmes are expensive, so if councils are to invest more efficiently in preventive care, they will need accurate and objective ways of determining risk at the individual level across their population.

In 2006, the Department for Communities and Local Government commissioned a study to explore whether it might be possible to build predictive models that identify people at risk of future admission to a care home. The initial report, *Predicting Who Will Need Costly Care*, was published by the King's Fund in November 2007.[3] It concluded that predictive tools are indeed theoretically possible, and that if a reliable predictive tool could be built then councils would be better placed to offer preventive interventions to the right vulnerable people and to construct robust 'business cases' for doing so.

Following the publication of the King's Fund report, the Care Services Efficiency Delivery programme at the Department of Health commissioned a team at the Nuffield Trust (many of whom had previously worked on the PARR/Combined Model project at the King's Fund) to test the feasibility of building predictive models for social care. The aim of this project, reported here, was to:

- obtain pseudonymous individual-level data from several primary care, secondary care and social care organisations
- link, collate and analyse these data at the individual level
- attempt to develop a statistical model to predict which individuals are at greatest risk of requiring intensive social care in the 12 months after prediction.

Box 1.1 Predictive risk models

PREDICTIVE RISK MODELS

Predictive risk models apply statistical techniques such as multiple regression or neural networks to analyse routine electronic data. They use historic patterns in the population's data to make predictions at the individual level. The growing use of predictive models in healthcare over recent years has been made possible by a combination of better access to individual-level electronic data and improvements in computing power. Datasets for large populations, often involving hundreds of millions of observations, can now be analysed according to individual health needs, service use and health outcomes.

In developing predictive models, it is crucial that they be 'generalisable', i.e. that they can be applied to datasets from other locations and timeframes. The standard approach for ensuring generalisability is to split the data at random, using half of the data (the 'development sample') to construct the model, with the other half (the 'validation sample') being used subsequently to test how well the model performs. The accuracy of each predictive model can be quantified using various measures based on its performance on the validation sample. These metrics include the sensitivity and specificity; the positive and negative predictive values; the area under the receiver operating characteristics curve (ROC curve); and the *r*-squared value.

In this report, we have concentrated on two of these measures:

- **Sensitivity**, which is a measure of how good the model is at detecting people from the population who will truly experience the outcome of interest (for example, admission to a care home).

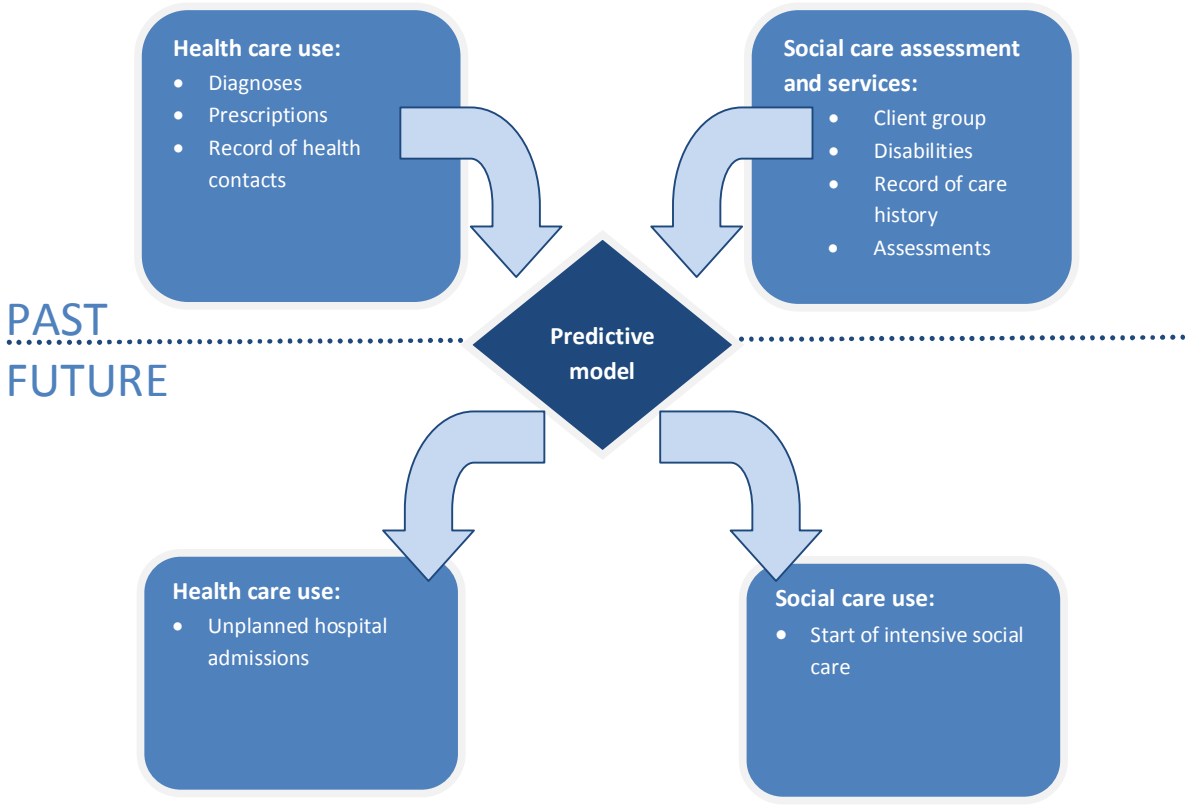
- **Positive predictive value (PPV)**, which is a measure of how reliable the predictions are that are made by the model, i.e. the chance that the people the model determines are at high risk of experiencing the outcome of interest (for example, admission to a care home) will indeed truly experience that outcome.

When a predictive model is used in practice, it is applied to the most recent data in order to produce a risk estimate for each individual in the population. Since the uncertainty surrounding the model's predictions is known from its performance on the validation sample, this can help commissioners to build robust business cases for early intervention.

We know from the literature that the predictor variables (or 'independent variables') for care home admission may include: age, sex, ethnicity, deprivation, morbidity, health service use, drugs prescribed, as well as patterns of social care needs and usage. Since these variables span health and social care records, a complicating yet critical prerequisite for this project was to link health and social care data at the individual level in a way that did not compromise confidentiality.

The focus of this project was on using patterns in historic healthcare data and social care data to predict future social care use. In addition, we tested the effect on predictive power of adding social care data to forecast future health service use (we found that it did increase predictive power, but only marginally).

Figure 1.1 Using health and social care data to predict health and social care usage



Predictive models for health care, such as PARR, use information about past health care use (top left box) to identify needs and then use these to predict future health care needs (bottom left box). To predict social care costs, we combined information on both health and social care use and needs (top two boxes) to predict future social care use (bottom right box).

Overall, the initial project involved six phases:

Initial project	
1 Approvals	Defining and obtaining all the necessary ethics and information governance approvals.
2 Data acquisition	Working closely with several paired primary care trusts and councils with social services responsibilities to work through the logistics of extracting, de-identifying, linking, encrypting and exporting their data (GP, in-patient, outpatient, accident and emergency (A&E), community, social services and housing data).
3 Descriptive analysis	Analysis of the data to describe patterns of social and healthcare use.
4 Modelling	Modelling social care costs and admissions to care homes using analogous methods to those employed in the PARR /Combined Model project for the NHS.[4]
5 Sensitivity analyses	Running many different variants of the models, and testing the use of additional datasets and classifications.
6 Business case tools	Developing tools that allow business cases to be developed spanning health and social care.

An important point to stress is that this work was a feasibility study. While none of the questions in Box 1.2 might be new individually, addressing all of these issues together in series was quite novel.

Box 1.2 Questions addressed by this feasibility study

- ***Is the information we need for modelling future social care outcomes stored in routine health and social care data?***
- ***Can we obtain appropriate agreements from the relevant parties to undertake this type of work?***
- ***Can we obtain individual-level information in sufficient quantities to permit modelling?***
- ***Can we link different datasets at a person-level in a way that does not compromise the confidentiality of the people using services?***
- ***Are the data recorded sufficiently accurately and completely to construct predictive models for social care?***
- ***Are the characteristics of users who will incur future high social care costs distinguishable within routine data?***
- ***Can we build a model that is statistically predictive of future social care use?***

2. DATA ACQUISITION

Evidence from case control studies and case note reviews published in the literature tells us that both health and social factors can be predictive of admission to a care home.[3] Therefore, for this project we required access to linked, individual-level information from both the NHS and local councils.

As with the Patients at Risk of Re-hospitalisation (PARR)/Combined Model project, the research team did not need access to any sensitive fields such as names, dates of birth or addresses.[4] Rather, we only required de-identified ('pseudonymous') data. In pseudonymous data, all of the sensitive fields have either been completely removed (for example, names) or truncated (date of birth truncated to age; address truncated to geographic area), and the unique key (in this case the NHS number) is transformed into a meaningless pseudonym. For the PARR/Combined Model project, all of the data came from the NHS and could therefore be linked using pseudonymous NHS numbers as the unique key. However, this project required the incorporation of council data, which do not routinely contain NHS numbers. Therefore, more sophisticated linkage techniques were required in some cases.

APPROVALS

Before we could apply to NHS organisations and councils for access to their data, we had to be sure that we had all of the necessary ethics and information governance approvals in place. We were obliged to negotiate these locally because there is currently no single overarching authority for these matters. Nor is there an unambiguous framework for sharing such data in the UK. This contrasts with the United States, where the Health Insurance Portability and Accountability Act (HIPAA) helps to clarify responsibilities.[5]

ETHICS APPROVALS

The UK's research ethics mechanisms have recently been streamlined into a National Research Ethics Service, which operates a single Integrated Research Application System (IRAS)[6]. Since our research only involved pseudonymous data, we were unable to submit an application through IRAS as it did not fall unambiguously into a category of research, as opposed to audit or service evaluation.[7] Instead, we applied to the local Research Ethics Committee (REC) for one of the five sites, which kindly provided written confirmation that ethics approval was not required for this study. IRAS subsequently provided us with an email confirming that the letter from the local REC could be applied nationally.

INFORMATION GOVERNANCE APPROVALS

This project involved the analysis of health and social care data at a person level. The government has made it clear that the fundamental principle governing the use of person-identifiable information by any part of the NHS or the research community is that of informed consent. However, the size of the datasets required for this project meant it would be unfeasible for us to seek individual consent from people to use their de-identified data for modelling. Normally, in situations where consent cannot be obtained, no information that identifies individual patients may be used. The only exception to this rule is for essential NHS activities that are in the interests of patients or the wider public, where anonymous or aggregated information will not suffice, and where obtaining consent is not a practicable alternative.[8] Under Section 60 of the Health and Social Care Act 2001, applications to use data in this way must be submitted to the Ethics and Confidentiality Committee (ECC)

of the National Information Governance Board for Health and Social Care. The predecessor of the ECC, which was called the Patient Information Advisory Group (PIAG), issued a ruling in 2006 that the above principles could be met by encrypting data in such a way that they became effectively pseudonymous.[9] In July 2008, we obtained written confirmation from PIAG that the processes we planned to use for this project would meet all of its requirements and that therefore no application was necessary under Section 60 of the Health and Social Care Act 2001.

The process of pseudonymisation involves the following steps:

- truncating the data (for example, converting date of birth into age and converting postcode into lower super-output area)
- removing person-identifiable fields (name, address and date of birth)
- replacing the NHS number with a meaningless pseudonym.

For this project, as well as rendering the data pseudonymous, we also ensured that all data were transferred and stored in encrypted storage media.

DATA REQUESTS AND TRANSFER

In order to minimise the time required for local data collection, projects of this type are specifically designed to exploit the data from operational systems – either collated data (such as Hospital Episode Statistics) or operational systems (such as extracts from electronic medical records in primary care). However, it is worth noting that there is an administrative and IT burden on the primary care trusts (PCTs) and councils who choose to participate in this type of work. Therefore, this project very much relied on the goodwill and support of PCT, council and care trust staff, for which we are most grateful.

We approached a number of PCTs and councils with social services responsibilities (CSSRs) to ask them if they wished jointly to volunteer to participate in this study. In order to take part, the sites had to have at least three years' worth of historic health and social care data available, and preferably five years. They had to be prepared to use a linkage key and a common encryption method for all datasets, and they needed the capability to extract the contents of local datasets in a suitable format.

Although there were several technical steps involved in the extraction of the data, we understand that these did not prove to be overly difficult. Extracting NHS data involved:

- writing a specific query to extract the particular variables that we required for this project
- truncating or re-formatting other fields (postcode, date of birth)
- excluding any person-identifiable fields (name, address, date of birth)
- setting time frames (start date and end date for the data to be extracted)
- reformatting the data into plain text files
- replacing the NHS number with a pseudonym.

For social care, councils typically held their data in a dozen or more tables that often could only be analysed using a graphical interface (by dragging icons). However, some councils were using the Tool for Rapid Analysis of Care Services (TRACS) developed by the Department of Health's Care Services Efficiency Delivery programme, which helped facilitate this process.[10, 11] A similar procedure was used for obtaining pseudonymous housing data. (Note: TRACS has now been superseded by the Tools for Rapid Integration of Public Submissions, TRIPS.[34])

The data systems used in social care were primarily developed to fulfil local operational needs and in general are not used directly for comparative analyses. Furthermore, there is currently no equivalent national dataset to Hospital Episode Statistics for social care. Consequently, different approaches to data collection were required for each CSSR involved in this project. For example, we found that some CSSRs have complex systems where a high proportion of the data fields are defined locally. These fields often incorporate free text, which can be difficult to analyse or incorporate in predictive models. This posed a particular challenge for developing pooled models for this project, where we required a set of common definitions variables that could be applied across the five sites. In order to maximise the number of person-years' worth of data available for analysis, we prioritised compatibility between the datasets of the different sites at the expense of a reduced number of variables. This meant that our modelling was restricted to a relatively small number of resource-intensive aspects of social care.

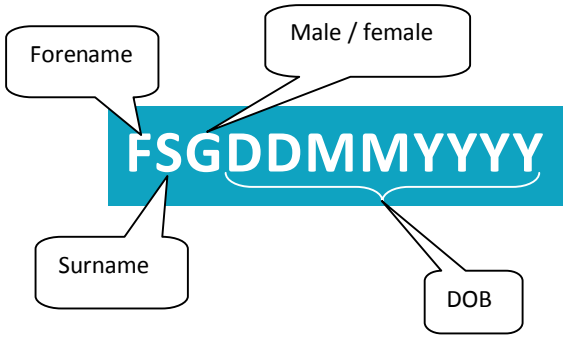
DATA LINKAGE

The factors predictive of care home admission are known to include both health and social care variables, so we needed to link health and social care data at the individual level for this project. This linkage had to be done in the absence of a unique key because at present social care data do not routinely incorporate NHS numbers. Data linkage also had to be performed in a way that did not compromise individual identities.

Our original intention had been to exploit NHS numbers in databases that are shared by a CSSR and a PCT, such as electronic Single Assessment Process (eSAP) databases. However, such databases were not used extensively in our sites, so we had to choose a different method. One approach we used successfully was to ask the PCT and CSSR to construct a new key using the first letter of the forename, first letter of the surname, the sex and the date of birth (see Fig 2.1). Using this method, we were able to achieve data linkage of approximately 90 per cent of records (see Section 3).

While it might be possible to improve the specificity of such a key, for example, by incorporating the first two letters of the surname rather than just the first letter, this also increases the opportunities for error (for example, McDonald versus Macdonald might now be wrongly rejected).

Figure 2.1 New key constructed from name, sex and date of birth



Housing data were linked in a similar fashion, but we were unable to link council tax because – in our sites at least – council tax data did not contain dates of birth.

An alternative way of linking health and social care data, which we have used successfully in a related project, is to establish a Batch Tracing Agreement[i] between the social services department and the NHS National Strategic Tracing Service, who then assign NHS numbers to the social services data.[12]

After data linkage was achieved, the sites truncated and/or stripped out the person-identifiable fields (name, address, date of birth) from their data. They then encrypted the NHS number before making the data available for us to collect. This ensured that we had no way of identifying any individual users at any stage of the project.

[i] In March 2009 the 'Batch Tracing Service' previously provided by the NHS Strategic Tracing Service was replaced by the 'Demographics Batch Service' of the new NHS Care Records Service (NHS CRS).

3. HEALTH DATASETS

HOSPITAL DATA

Information about NHS hospital activity has been collected in a standard format for many years. We used three distinct hospital datasets in this project: inpatient, outpatient and A&E. Each of the five sites was requested to provide three years' worth of data from all three of these datasets. Table 3.1 summarises the data we received.

Table 3.1 Numbers of records in the hospital datasets and the numbers of unique patients

		Site A	Site B	Site C	Site D	Site E
Inpatient	Time covered	**April 2006 – Jan 2009**	**April 2004 – Nov 2008**	**April 2005 – Aug 2008**	**April 2003 – Aug 2008**	**April 2005 – April 2008**
	No. records (episodes)	119,176	118,154	1,394,375	530,874	1,105,860
	No. spells	100,107	103,102	1,181,104	468,798	978,679
	No. patients	44,649	34,744	388,697	180,444	350,891
Outpatient	Time covered	**April 2006 – Jan 2009**	**April 2004 – Nov 2008**	**April 2003 – Aug 2008**	**April 2003 – Aug 2008**	**April 2005 – April 2008**
	No. records (attendances)	567,297	455,574	4,946,698	2,826,165	4,018,910
	No. attendances	523,30	444,37	4,078,426	1,956,357	3,058,678
	No. patients	86,134	55,604	559,828	273,798	556,996
A&E	Time covered	**April 2006 – Jan 2009**	**April 2004 – Nov 2008**	**April 2005 – Aug 2008**	**April 2003 – Aug 2008**	**April 2005 – April 2008**
	No. records (visits)	65,181	140,84	627,894	739,251	694,521
	No. patients	33,473	63,482	280,649	230,144	316,949

For predictive modelling, the task was to move from these 'event-based' data to new datasets that summarised events longitudinally at a person level. In other words, we needed to generate a single row of data for each person that contained a record of all pertinent events (both predictor events and outcome events) across a standard period of time (Year 1, Year 2 and Year 3).

Hospital data were first processed in order to construct a series of specific variables that would be used in the modelling phase. Table 3.2 summarises the predictor variables that we created from two years' worth of inpatient data (Year 1 and Year 2).

Table 3.2 Summary of variable types derived from hospital inpatient data

Variable group	Notes	Time period(s)
Number of emergency admissions		prior 1–90; 91–180; 181–365; 366–730 days
Number of ordinary elective admissions		prior 1–90; 91–180; 181–365; 366–730 days
Number of day case admissions		prior 1–90; 91–180; 181–365; 366–730 days
Number of emergency 'avoidable' admissions	Based on a list of ambulatory care sensitive conditions (see below) that are defined using Healthcare Resource Groups (HRG).	prior 1–90; 91–180; 181–365; 366–730 days
Number of emergency 'medical' admissions	HRG derived	prior 1–90; 91–180; 181–365; 366–730 days
Number of emergency 'mental health' admissions	HRG derived	prior 1–90; 91–180; 181–365; 366–730 days
Number of emergency 'alcohol and drug' admissions	HRG derived	prior 1–90; 91–180; 181–365; 366–730 days
Number of emergency 'cancer' admissions	HRG derived	prior 1–90; 91–180; 181–365; 366–730 days
Number of admissions where the patient ultimately self-discharged		prior 1–365; 366–730 days
Long-term condition groupings (27 variables)	A list of 27 common conditions. A flag for each was created based on the presence of a relevant diagnosis in patient history. Derived with reference to diagnosis codes.	prior 1–90; 91–180; 181–365; 366–730 days
International Statistical Classification of Diseases and Related Health Problems (ICD) care groups derived disease groupings (21 variables)	In-house classification of diagnoses felt to be associated with frail elderly. Derived with reference to diagnosis codes.	Prior 1–730 days
Hierarchical condition categories (HCCs) (70 variables and 6 interaction terms)	HCC definitions derived from published sources. Derived with reference to diagnosis codes.	Prior 1–730 days
Average number of episodes per spell	Elective and emergency separate. Acts as a proxy measure of complexity.	Prior 1–365; 366–730 days

As can be seen from Table 3.2, as well as summarising activity by type (elective, emergency, day case), it was necessary to try a number of different approaches to organising and classifying the wealth of diagnostic information available from inpatient data. The classification techniques we used to describe hospital admissions included hierarchical condition categories (HCC)[13]; Adjusted Clinical Groups™ (ACG); long-term conditions (LTC); ambulatory care sensitive conditions (ACS) and diagnostic classifications.

Table 3.3 Hospital admissions* for people aged 75+

	Site A	Site B	Site C	Site D	Site E
Number of admissions in people with no prior 'significant' social care use	15,658	9,093	76,809	23,202	53,600
No. emergency admissions per 1,000	301	271	277	403	366
No. ordinary elective admissions per 1,000	107	118	166	132	89
No. day case admissions per 1,000	227	311	331	282	248
No. emergency 'avoidable' admissions per 1,000	86	79	78	118	104
No. emergency 'medical' admissions per 1,000	225	202	208	314	279

*Note that the values for patients per 1,000 head of population aged 75+ were only used as a crude check, and do not adjust for the differences in the age and sex profiles of the sites.

In contrast to inpatient data, outpatient data contain very little information about patients' diagnoses. Outpatient data record the specialty concerned, but this classification is often very broad (for example, 'general medicine' or 'general surgery'). The number of different specialties in a person's history was used as a proxy for potentially complex co-morbidities. Otherwise, apart from specialty, we could only rely on the numbers and types of attendances over a given period to provide further information. Table 3.4 details the outpatient variables that created for every person.

Table 3.4 Summary of variable types derived from hospital outpatient data

Variable set	Time period(s)
Total number of attendances	Prior 1–90; 91–180; 181–365; 366–730 days
Number of urgent attendances	Prior 1–90; 91–180; 181–365; 366–730 days
Number of referrals initiated by consultant responsible for outpatient (OP) attendance	Prior 1–90; 91–180; 181–365; 366–730 days
Number of referrals initiated by consultant other than in an A&E dept (and not responsible for OP attendance)	Prior 1–90; 91–180; 181–365; 366–730 days
Number of referrals from GP	Prior 1–90; 91–180; 181–365; 366–730 days
Number of referrals from A&E	Prior 1–90; 91–180; 181–365; 366–730 days
Number of referrals from other sources	Prior 1–90; 91–180; 181–365; 366–730 days
Number of referrals where another appointment was made	Prior 1–90; 91–180; 181–365; 366–730 days
Number of referrals cancelled by the patient or not attended	Prior 1–90; 91–180; 181–365; 366–730 days
Speciality of attendances (45 variables)	Prior 1–730 days
Number of different specialties	Prior 1–730 days

A&E data have only recently been incorporated within Hospital Episode Statistics (HES). These data contain information about the arrival of patients to A&E, any investigations performed, as well as the diagnoses made. The data are known to be of poor quality nationally (just over 60 per cent of activity was captured in the 2007/08 collection of national HES A&E data[14]). But as well as problems of incompleteness, the codes in the dataset (diagnostic codes, for example) are used inconsistently both within and between sites. Despite these problems, A&E data do offer some meaningful insights into the experiences of people who are at risk of requiring intensive social care. The variables we created from the A&E data are shown in Table 3.5.

Table 3.5 Summary of variable types derived from hospital outpatient data

Variable set	Time period(s)
Total number of attendances	Prior 1–90; 91–180; 181–365; 366–730 days
Number of urgent attendances	Prior 1–90; 91–180; 181–365; 366–730 days
Number of referrals initiated by the consultant responsible for an outpatient attendance	Prior 1–90; 91–180; 181–365; 366–730 days
Number of referrals initiated by a consultant other than in an A&E department (and not responsible for an outpatient attendance)	Prior 1–90; 91–180; 181–365; 366–730 days
Number of referrals from GP	Prior 1–90; 91–180; 181–365; 366–730 days
Number of referrals from A&E	Prior 1–90; 91–180; 181–365; 366–730 days
Number of referrals from other sources	Prior 1–90; 91–180; 181–365; 366–730 days
Number of referrals where another appointment was made	Prior 1–90; 91–180; 181–365; 366–730 days
Number of referrals cancelled by patient or not attended	Prior 1–90; 91–180; 181–365; 366–730 days
Speciality of attendances (45 variables)	Prior 1–730 days
Number of different specialties	Prior 1–730 days

GP DATA

Models that incorporate GP encounters data (such as the Combined Model) do not necessarily require an initial hospital admission as a 'triggering event'. This is in contrast to predictive models – such as the Patients at Risk of Re-hospitalisation (PARR) model – that only use hospital data, and which can therefore only make predictions on people who have already been to hospital in the preceding years.

GP data were used for two reasons: first, for determining the registered population at each site and, second, for the clinical information held in the GP electronic medical record. Such clinical information was only available from two of the five sites we studied.

GP REGISTRATION DATA

Each primary care trust (PCT) or care trust provided a GP registration file, also known as an 'Exeter file'. This was used as the core population file onto which all other data were attached for modelling. The Exeter file represents a snapshot at a point in time; for some sites we obtained more than one such snapshot. Having an entire population file is important for modelling because those people registered with a GP, but otherwise not in contact with healthcare services, may still have been receiving social care services. It is also important to include in the model those people with no use of any services at all, so as to determine the model's coefficients correctly.

Table 3.6 shows the details of the population data we received from the four PCTs and the care trust. While we ultimately used only a single year's population data for modelling (that of April 2007 in all cases except for Site B) we, in fact, received as many as six years' data from some sites. The initial models were based on the population of people aged 55 and above, but later we restricted ourselves to the population aged 75 and over, for reasons discussed later. As can be seen from Table 3.6, there was an approximately eight-fold difference between the population of the biggest area and that of the smallest. In each site, about 60 per cent of people aged 75 and over were female. It seems that the number of deaths in the prediction year (Year 3) in Site B was much lower than that observed in the other sites.

Table 3.6 GP registration files ('Exeter Files') received from the five sites

	Site A	Site B	Site C	Site D	Site E
Dates of datasets provided	Yearly: April 2003 to April 2008 inclusive	April 2008	Yearly: April 2003 to April 2008 inclusive	Yearly: Jan 2004, April 2005 to April 2008 inclusive	Yearly: April 2005, April 2006, April 2007
Total number of people-years (years)	864,170 (6 years)	100,521 (1 year)	4,395,509 (6 years)	1,911,693 (5 years)	589,431 (1 year–2007)
Selected time period for modelling	April 2007	N/A	April 2007	April 2007	April 2007
Practice coverage	All practices (21)	18 practices (out of 38)	All practices (108)	All practices (66) plus residents[ii]	All practices (57) plus residents
Number people	145,027	(100,521)	741,290	374,494	589,431
Number aged 75 plus	16,839	9,512	82,847	23,983	53775
75 plus: percentage females	61.5%	60.4%	60.4%	60.6%	61.1%

[ii] People who live in the geographic boundaries of the PCT but are not registered with a GP practice within those boundaries (six per cent of people).

GP ENCOUNTERS DATA

One of the intentions of this study was to see if the clinical detail within records of GP encounters would help identify specific characteristics of older people that were predictive of the future use of intensive social care services: i.e. would this information create a more accurate predictive model. Therefore, in addition to the basic population file that we obtained from all of the sites, in two of the sites (Sites D and B) we were able to obtain detailed information on people's GP encounters. These GP encounter data are potentially extremely useful because they contain very rich clinical information; however, this also makes them challenging to obtain and to analyse. For predictive models of hospital admission, the addition of GP information allowed a predictive model to be built that can identify a much wider section of the population at risk (Combined Model versus PARR[15]).

The extracted GP encounter data received from the two sites had a relatively simple structure:

- pseudonymous patient identifier
- date
- coded thesaurus of clinical terms, known as Read Codes[iii]
- up to two fields to record pertinent values (values of any tests or blood pressure readings, for example).

Any one visit to the GP may result in several different Read Codes being recorded. For example, in one site the data contained over 82 million records that were coded using over 57,000 different Read Codes.

The aim of analysing the GP encounters data was to identify those Read Codes that might indicate one of the health problems that we know from the literature can be predictive of future social care use. Unfortunately, the complexity and non-hierarchical nature of Read Codes meant that this was not simple. Three different approaches were used:

- a pragmatic approach looking at a limited set of health conditions whose codes occurred relatively frequently
- selecting variables used in the creation of the Combined Model to predict hospital admission
- a subset of groups derived from the Adjusted Clinical Groups™ (ACGs) system developed at Johns Hopkins University.[16]

[iii] See www.connectingforhealth.nhs.uk/systemsandservices/data/uktc/readcodes

Table 3.7 Common Read Codes

	Site B		Site D		Both sites	
GROUPING	No. mapped Read Codes	Count of events	No. mapped Read Codes	Count of events	Count of events	% of all
No group	40082	26,191,161	56071	75,678,066	101,869,227	91.8%
Diabetes	190	280,303	231	1,406,544	1,686,847	1.52%
Hypertension	19	347,218	19	1,191,354	1,538,572	1.39%
Asthma/COPD	9	236,515	11	972,094	1,208,609	1.09%
Depression	17	285,926	17	575,126	861,052	0.78%
Heart failure or heart disease	5	164,538	5	522,070	686,608	0.62%
Heart disease/angina	9	99,972	9	321,581	421,553	0.38%
Anxiety (tranquilisers)	69	123,306	74	290,033	413,339	0.37%
Malnutrition	153	108,886	252	241,541	350,427	0.32%
Osteoporosis	7	118,191	9	189,284	307,475	0.28%
Psychosis	138	82,483	199	205,707	288,190	0.26%
Atrial fibrillation	3	71,646	3	212,128	283,774	0.26%
Anaemia	3	52,810	3	131,393	184,203	0.17%
Urinary incontinence	57	45,514	69	96,508	142,022	0.13%
Parkinson's disease	73	31,151	82	101,079	132,230	0.12%
Mental health	7	19,071	12	75,221	94,292	0.08%
Home visits	16	7,216	20	74,857	82,073	0.07%
Glaucoma	6	23,034	6	57,419	80,453	0.07%
Obesity	18	25,627	17	51,672	77,299	0.07%
Stroke	15	17,023	14	41,015	58,038	0.05%
Mobility	13	9,274	15	38,209	47,483	0.04%
COPD	15	19,116	12	16,651	35,767	0.03%
Autoimmune disease	1	15,021	1	14,937	29,958	0.03%
Falls	14	9,611	17	17,890	27,501	0.02%
Neurological disease	1	4,875	1	13,710	18,585	0.02%
Social care service indication	7	1,525	8	10,295	11,820	0.01%
Dementia	35	4,171	32	6,564	10,735	0.01%
Dependence (on carer/other)	4	1,797	4	7,237	9,034	0.01%
Isolation	3	1,180	3	4,518	5,698	0.01%
Nursing home/other	5	1,648	5	3,069	4,717	0.00%
Bowel Incontinence	5	820	8	1,468	2,288	0.00%
Confusion	6	1,451	4	828	2,279	0.00%
Dehydration	2	131	2	272	403	0.00%
Blindness/deficiencies of vision	1	85	1	243	328	0.00%

4. SOCIAL CARE AND OTHER DATASETS

The conventions and classifications in place for social care data are much less standardised than those for healthcare data. Although social care data tend to be structured differently in each council with social services responsibilities (CSSR), we were able to obtain three types of information from all five sites, namely demographics, assessments and utilisation data.

In order to deal with the large diversity of services recorded in each area, we opted to group services into broad categories (see Box 4.1). Grouping services together in this way has a number of advantages: it creates more consistency between the sites; it simplifies the dataset used in the models; and it allows us to apply national unit costs. However, the downside is that grouping relies on judgements being made about the nature of each service for which only limited information was available from the operational systems.

Box 4.1 Grouping social care services

Home care

This would ideally be defined in line with Home Help / Home Care (HH1) guidance[17] to include:

- traditional home help services (including home help services provided by volunteers)
- overnight, live-in and 24-hour services
- practical services which assist the client to function as independently as possible and/or continue to live in their own homes (for example, routine household tasks within or outside the home, personal care of the client, shopping, overnight, live-in and 24-hour services, respite care in support of the client's regular carers).

Residential care

This would ideally be defined in line with Supported Residents Collection (SR1) guidance[18] as accommodation with both board and personal care for persons requiring personal care by reasons of disablement, past or present, dependence on alcohol or drugs, or past or present mental disorder.

Nursing care

This would ideally be defined in line with SR1 guidance to include nursing and other medical care provided in premises defined in Sections 21 to 22 of the Registered Homes Act 1984.

Residential respite care

This relates to help and support that allows an individual to take a break from the responsibility of caring for somebody else. It does not include day care or home care.

Other accommodation

This includes sheltered housing, very sheltered housing and extra care housing, as well as unstaffed (group) homes.

Equipment and adaptations

Note: telecare has been classified as equipment rather than home care.

Direct payments

These are cash payments made to individuals who have been assessed as needing services and are in lieu of social service provisions.

Day care

This is usually, although not always, offered in day centres. It includes services designed to assist people in maintaining links with the community and in avoiding social isolation. It can also provide carers with an opportunity to have their own space and time. It includes transport to and from day care.

Meals

Other

Not classified above. This may include counselling, training, etc.

Home care was subdivided into high-, medium- and low-intensity services, based on the number of contact hours per week recorded. Low-intensity home care was defined as less than two hours a week of care; medium-intensity home care was two to ten hours; and high-intensity home care was more than ten hours a week. This is similar to the classification used in the HH1 returns collected by the NHS Information Centre. Overnight services were classified as high-intensity regardless of the number of contact hours recorded on the data system. Table 4.1 summarises the number of people by service type for one year across the five sites.

Table 4.1 Number of person-years in receipt of social care in prediction year for people aged 55+ (absolute numbers rather than rates per 1,000 population)

	Site A	Site B	Site C	Site D	Site E	All sites
Meals	335	819	1,283	322	331	3090
Day care	499	1346	1,990	200	181	4217
Home care	1,399	5,462	3,929	1,095	5,525	17,409
Res. care	788	972	3,481	285	126	5,652
Nursing care	498	943	2,142	220	156	3,959
Direct payment	127	58	606	18	0	808
Other accommodation	0	82	0	275	171	528
Respite	0	367	18	32	0	418
Other	432	1611	383	677	92	3,195

Costs were estimated by applying published national unit costs to the social care utilisation observed in order to calculate weighted utilisation. Therefore, where the rest of this report refers to the 'costs' of social care episodes it should be noted that we have not performed full economic costing. Another point to note is that there is a choice between applying national average unit costs or unit costs for the individual CSSRs. We chose to use a national average unit cost in order to allow us to make comparisons between the sites.

In this report, we have not distinguished between services provided by a CSSR and those provided by private contractors on behalf of the CSSR. This is because to do so would have required us to make additional judgements about the coding systems used by the councils. In reality, unit costs are substantially *lower* for privately provided services than they are for services provided by CSSRs themselves.

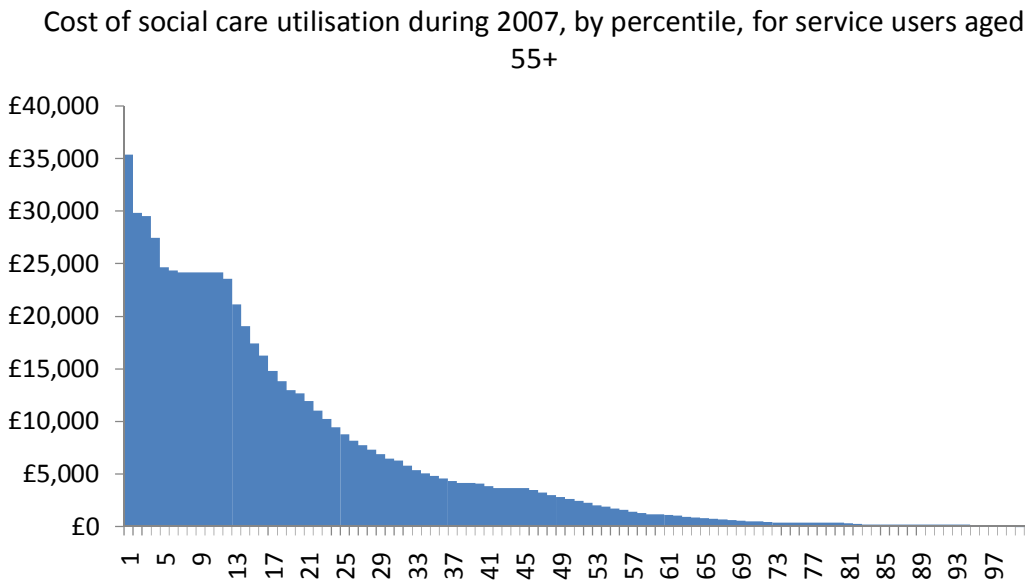
Unit costs are calculated according to the gross cost of providing the service before any user charges have been deducted. They therefore reflect the combined financial impact of these services on the CSSR and the client. The unit costs used are set out in Table 4.2. Some technical improvements in the accuracy of these costs may be possible. For example, we could have explored whether it was possible to extract the data on the monetary amount of direct payments from the operational systems; whether it was possible to distinguish equipment from adaptations; and whether it was possible to introduce a classification of day care into high- and low-intensity packages. Some of the participating CSSRs advised us that certain adaptations, direct payments and day care can be very expensive. However, with the unit costs we used, the highest intensity services were nursing homes, residential homes and high-intensity home care.

Table 4.2 Unit cost assumptions

Service group	Unit cost	Source
Nursing care	£568 per week	National average across all CSSR and other provision for older people from PSSEX1 2007/08, plus an allowance for the NHS contribution to nursing care in nursing homes
Residential care	£465 per week	National average across all CSSR and other provision for older people from PSSEX1 2007/08
Home care:		Based on average hours received per group and an assumed cost of £15.20 per hour (national average across all CSSR and other provision for adults and older people: PSSEX1 2007/08)
high-intensity	£244 per week	
medium-intensity	£71 per week	
low-intensity	£16 per week	
Respite	£465 per week	Assumed same as residential care
Other accommodation	£465 per week	Assumed same as residential care
Equipment and adaptations	£176 per installation	Calculation based on PSSEX1 and RAP
Direct payments	£124 per week	National average for older people from PSSEX1 2007/08
Day care	£80 per week	National average for older people from PSSEX1 2007/08
Meals	£22 per week	National average for older people from PSSEX1 2007/08
Other	Not costed	

As expected, the cost of social care provision was very unevenly distributed across the population. People in the top quintile of service users cost around £35,000 each a year, compared with a mean cost of around £5,500 each a year (Figure 4.1). There was a concentration of service users whose cost is around £24,000 a year, corresponding to a year in a residential home. Note that even if someone received nursing care for a full year, this would cost £29,500 – so the most expensive service users were recorded as receiving more than one type of service. Overall, high-intensity services (nursing homes, residential homes and high-intensity home care) accounted for around 70 per cent of a CSSR's observed expenditure on social care.

Figure 4.1 Distribution of social care expenditure across the population



The other type of information held in social care operational systems are the 'assessment' files that contain information on people's social care needs (see Table 4.5). Some, but not all of the participating CSSRs provided us with information on Fair Access to Care (FACS) bands, which is a framework for determining eligibility for

adult social care. In addition, we were able to identify some very simple indicators relating to hearing or visual impairment. These were recorded in the data received from four of the five CSSRs.

- Social isolation – defined as living alone, being widowed or separated, or where the objective of the service was to promote social inclusion.
- Functional limitation – recorded as difficulties with activities of daily living (ADLs) (such as personal care, walking, bathing or dressing) or with instrumental activities of daily living (IADLs) (such as meal preparation, domestic tasks or shopping).
- Health condition – including dementia, depression and incontinence.

Table 4.3 Definitions of needs variables contained in social care data[iv]

	Impairment	Social isolation	Functional limitation	Health condition	Presence of informal carer
Site A	Hearing or visual impairment	Living alone	History of problems	Mental health condition or incontinence	Living with family or spouse
Site B	Hearing or visual impairment	Living alone, single, or outcome	Outcome-based measures	None	Sign of carer need
Site C	Hearing or visual impairment	None	None	Dementia	Sign of carer need
Site D	Hearing or visual impairment	Single or outcome	List of ADLs or IADLs	General marker of other health	Sign of carer need or carer recorded

We also noted any flags in the data indicating the presence of an informal carer. There were differences between the five sites in the way such information was recorded, but it is important to stress that any differences we observed only reflected what was recorded in each site, not any underlying phenomena. With a few exceptions, it was found that the social care needs tended to be recorded more frequently among people who went on to receive a service rather than among people who did not go on to receive a service.

Table 4.4 Number of person-years in receipt of social care in 2007 for people aged 55 and over (absolute numbers rather than rates per 1,000 population)

	Site A	Site B	Site C	Site D	Site E	All sites
Meals	335	819	1,283	322	331	3090
Day care	499	1346	1,990	200	181	4217
Home care	1,399	5,462	3,929	1,095	5,525	17,409
Residential care	788	972	3,481	285	126	5,652
Nursing care	498	943	2,142	220	156	3,959
Direct payment	127	58	606	18	0	808
Other accommodation	0	82	0	275	171	528
Respite	0	367	18	32	0	418
Other	432	1611	383	677	92	3,195

[iv] For Site E we did not obtain any linked social care needs variables.

Table 4.5 Recorded needs for people with an assessment recorded on social care data (Note: this is not restricted to any particular time period or age group)[v]

	Impairment	Social isolation	Functional limitation	Health condition	Presence of informal carer
People with an assessment – did not go on to receive a service					
Site A	3%	5%	15%	7%	7%
Site B	7%	30%	16%	0%	6%
Site C	8%	0%	0%	1%	1%
Site D	4%	13%	3%	0%	5%
People with an assessment – went on to receive a service					
Site A	3%	7%	51%	37%	7%
Site B	9%	49%	12%	0%	6%
Site C	6%	0%	0%	5%	0%
Site D	12%	36%	40%	0%	22%

Within social care data, service users may have health or social care needs recorded, although many people receiving social care do not have any needs recorded. Users with a recorded social care need recorded in social care data cost less, on average, than service users as a whole (£5,921 in comparison to £6,465); whereas people with a health need recorded in social care data cost more, on average (£10,300) than service users as a whole (see Table 4.6). However, there is a wide variation of costs within each needs classification. Interestingly, the needs classification only explains around two per cent of the total variation of costs among social care users, so most of the variation is being driven by differences that are not recorded in the data.

Table 4.6 Social care users in 2007 aged 55 and over by needs recorded during 2005 or 2006

	Proportion of service users with this type of need recorded	Mean cost (£) of social care services in 2007	Standard deviation of cost (£) of social care services in 2007
Impairment	8%	5,478	7,804
Isolation	22%	5,106	7,307
Presence of informal carer	6%	5,381	7,888
Functional limitation	17%	5,792	8,775
Health problem	8%	10,300	11,148
Two or more needs	15%	6,396	9,004
Any of these needs	43%	5,921	8,453
None of these needs	57%	6,880	8,984
All users	100%	6,465	8,771

[v] For Site E we did not obtain any linked social care assessment variables.

OTHER DATA

As well as hospital, GP and social care data, there are many other routinely available data sources that contain variables known from the literature to be predictive of care home admission. We explored the feasibility of extracting housing data; community health services data such as district nursing; and council tax data. A range of other potentially valuable datasets were considered (such as Ambulance Trust data, Mental Health Trust data, data from the Third Sector and data on bereavement), however, it was not practicable to obtain data within a reasonable time frame. In many cases it would have required negotiating with a new set of organisations, or (as in the case of bereavement data) it would have involved very complex data linkage.

HOUSING DATA

Housing data may relate to:

- applications for housing (the housing register)
- housing needs and assessments
- supply of housing arranged for or provided by the council
- sheltered housing
- Supporting People

In one of our sites, housing data were held at district rather than county level and it was not judged practicable to negotiate with a number of district councils. Therefore we only had housing data from two sites where we looked at distinct sources of information. The detail of information on housing registers is limited. In one site only cross-sectional data were available, relating to the housing register at one particular point in time. Ideally, we would have received longitudinal data of the housing register over at least a two-year period. The result was that the dataset was very small (around 240 records) and so it was not suitable for predictive modelling. In another site we found that longitudinal data were available, but only for one particular locality.

Other forms of housing data were either found to be held by third party organisations rather than by the council (such as for clients of the Supporting People programme or residents of social housing provided by housing associations and registered social landlords), or could not be accessed in time for this project.

Throughout this process, our data collection efforts were hampered by the need to clarify arrangements for information governance. The Ethics and Confidentiality Committee (ECC) confirmed that housing data would not be patient data in terms of the ECC's remit, other than, for example, where the place of residence could be used to infer health information, such as a mental health institution. This made collection of housing data more problematic than health and social care information, where there is the ECC to confirm that the use of pseudonymous data does not require approval under section 251 of the NHS Act 2006.

COMMUNITY HEALTH SERVICES DATA

Community health services data relate to those healthcare services that are provided directly by a primary care trust (PCT) or by a local community services provider in the community. These include district nursing, community physiotherapy, podiatry, etc. In three of our five sites, community services data were paper-based and so were, therefore, unavailable for modelling (since electronic databases are required).

We received a large set of community health services dataset from one site which included 1.3 million community services contacts for 55,000 patients over a period of five and half years. Encrypted NHS numbers were available only for people who were registered in the PCT as of November 2008, so it was only possible to

link a subset of cases. Nevertheless, this did cover over 900,000 contacts with community services for 41,000 patients.

There were significant overlaps between people who used community health services and social care. Overall, 31 per cent of patients receiving community health services were also in receipt of social care, but this rose to 67 per cent for patients receiving community dentistry.

Table 4.7 Overlap between community services and social care in Year 2 of the prediction period (September 2006 to August 2007) for people aged 75+

	Number of community services patients	Number of patients also receiving social care service	Proportion of patients also receiving social care service
District nurse	2,636	1,252	47%
Chiropodist/podiatry	3,965	1,086	27%
Physiotherapist	1,207	429	36%
Health visitor	493	250	51%
Occupational therapist	484	235	49%
Support worker	279	152	54%
Speech and language therapist	241	101	42%
Health visitor for older people	125	65	52%
Community dentist	84	56	67%
Community matron	100	54	54%
Any community service	**6,953**	**2,158**	**31%**

COUNCIL TAX DATA

Council tax data are potentially useful because they contain markers of living alone and disability and a proxy measure of housing wealth. As with housing data, we knew that it was unfeasible to seek data from one of the five sites because council tax data there are held at the district level rather than the county level, so this would have involved negotiating with several organisations.

We explored the possibility of linking council tax data from two other sites. In both, we found that the information that would be required to link the data with health and social care data was not available on the council tax datasets. Thus it was not possible to obtain the NHS number, date of birth, or sex of council tax payers – information required for our various data-linking strategies. Further, information was typically available only for the person who pays the tax and not necessarily for other adults in the household. Unfortunately, it was not possible for us to obtain and link council tax data for this project.

MOSAIC™ SOCIOECONOMIC VARIABLES

Mosaic™ UK is the latest version of a consumer classification system developed by Experian UK.[19] It covers the whole of the UK and classifies every person or household into 61 types that can be aggregated into eleven groups. Using over 400 data variables and updated annually, it paints a rich picture in terms of demographics, socioeconomics, lifestyles, culture and behaviour. Experian provided us with a research licence to test the additional predictive power of including a subset of the Mosaic™ variables within our models.

It was important to keep the number of additional variables to a minimum to avoid over-fitting, since our datasets did not have huge numbers of cases. The data were first grouped up from Census Output Area level to Lower Super Output Area (LSOA). For all of the variables of interest, each of the 32,000 LSOAs was placed into its appropriate quintile. Quintile 1 related to the 20 per cent of LSOAs with the 'lowest' variable values, and quintile five to the 20 per cent with the 'highest' values. The quintile scores were then added to our modelling file for a single site (Site D). The intention was to replace each of these five-point ordinal variables with a binary variable. For each variable in turn, the relative proportion of people receiving a significant social care service or having an increase of costs of £1,000 was studied for each of the five quintiles. A subset of quintiles was selected where there appeared to be sufficient variation to offer some prospect of these variables adding discrimination (see Table 4.8).

Table 4.8 Selected Mosaic™ variable quintiles and associated significant social care use (Site D)

Mosaic™ variables created		Quintiles selected (5 = highest)	Proportion of people with a significant new service or £1,000 increase in costs	
Proportion of households: one person		4 & 5	1.7%	1.0%
Proportion of households: single room		5	1.7%	1.0%
Proportion with at least one county court judgment		4 & 5	1.7%	1.1%
Estimated proportion with two or more county court judgments		4 & 5	1.7%	1.1%
Estimated proportion of population who are heavy smokers		4 & 5	1.8%	1.2%
Estimated proportion of households: single person		5	1.2%	1.7%
Estimated proportion of households: owner occupied		1	1.9%	1.1%

5. BUILDING A PREDICTIVE MODEL

Predictive risk modelling is a technique that assesses the strength of relationships between variables and uses these relationships to forecast future events. For this project we tested whether we could build a model using variables that describe a person's health and social condition at one point in time in order to predict a subsequent change in that person's use of social care. This section outlines the approach taken and presents the summary results for a series of 'base' models. Later sections describe several different variants to the base models presented here.

TIMESCALES

The timescales for the datasets used were slightly different for each of the five sites, so we had to assume that there were no significant seasonal effects within the data. All models were based on predicting over a period of one year. This has the advantage of simplicity but it does assume that the relationships between the needs variable and a subsequent service change operate on similar timescales. Ideally, we would have wanted to look at smaller time periods – for example, whether health events in Quarter 1 led to a change in social care use in Quarter 2. However, the numbers of events we were trying to predict was already relatively small and using such an approach makes modelling increasingly difficult as sample sizes dwindle.

ANALYSIS OF NATIONAL HOSPITAL DATASETS

We undertook some work – using national datasets (Hospital Episode Statistics (HES)) – to see if the patterns of hospital admission and discharge codes could be used for modelling. More specifically, we tested whether diagnostic information and data relating to inpatient utilisation were predictive of a future hospital spell that ended in discharge direct to a care home. This analysis was conducted on a subset of 20 primary care trusts (PCTs) with a total of over a million records. In short, the results were disappointing. There are several possible explanations:

- We were trying to predict two events in a series (hospital admission and then discharge destination), which may have been too complex.
- Health data alone may be insufficiently predictive of care home admission.
- The recording of discharge destinations within HES may be too unreliable.

We discounted the first of these hypotheses by building a model designed to be run at the time of admission to hospital that used NHS data alone to predict the discharge destination from hospital. We found that this model still produced unreliable predictions.

We studied the reliability of discharge destination in detail using local linked social care and inpatient data from the secondary uses service (SUS) from which HES data are derived. This showed a high level of inconsistency between the recording of admissions to social care on the SUS record and that on the social care record. Although the results of this exercise were disappointing, it did at least demonstrate the need to use social care data for making predictions of future social care use.

LOCAL MODELS

The first step was to randomly split all of the data we received from each site. Half of the data were used to develop the models (the 'development sample') and the other half were used to test the accuracy of our newly developed models (the 'validation sample'). This 'split test approach' is a standard practice to counter the problem of 'over fitting', in other words, of developing a model where the relationships between the variables are overly dependent on the characteristics of the dataset under study, and where the results would therefore be less reliable if applied elsewhere (that is, poor generalisability).

The basic model aimed to predict whether a person had a significant change in social care use (as defined below) in the target year. It made predictions for Year 3 based on information drawn from Years 1 and 2. The independent (predictor) variables from Years 1 and 2 that we included in this model were of the following types:

- demographic characteristics
- health needs/problems
- social care needs/problems
- use of health services
- use of social care services
- socio-demographic variables pertaining to the area of residence.

Logistic regression was used to predict a binary event, namely a change in social care usage in Year 3. The regression produced a score between zero and one for each person, which was multiplied by 100 to give a score between zero and 100. This score represented the likelihood that the person would begin intensive social care (see below) in Year 3.

In this report, the performance of the models is shown using a threshold score of 50; any individuals with a score greater than 50 are deemed by the model as being 'at risk'. This threshold can, however, be adjusted. A lower threshold will yield a greater sensitivity but a lower positive predictive value, and vice versa.

POOLED MODELS

As well as developing local models, we also examined the effects of creating pooled models that combined information from a number of different sites. These pooled models used a set of predictor variables drawn from local models, rather than establishing a completely new set of variables.

An advantage of pooled models was that the results would be applicable to a number of sites – and so implicitly these models have a greater degree of generalisability. The larger numbers of events in the pooled model also help build more robust conclusions. However, the disadvantage of pooled models is that if there are important differences between the definitions of variables between sites, or in the relationships between the explanatory variables recorded, then they will tend to increase the variability of the results.

INDEPENDENT (PREDICTOR) VARIABLES

The basic prediction models started with the whole population registered with GPs within each of the five sites. This meant that the whole registered population was included, not just those people who had some contact with healthcare or social care services in Years 1 and 2. As a result, a large proportion of the records were not relevant to analysis of social care services for older people. Therefore the analyses were restricted to the older population, initially aged 55 and above, but later narrowed to people aged 75 and over. This focused the modelling process and meant that there was a higher chance of social care use in our study datasets.

Information about prior use of health and social care services was included as proxies for health or social care needs that were not otherwise identifiable from the data. For example, it was assumed that a previous visit to an A&E department signified some form of health problem. Likewise, we used the fact that a social care service provided in a previous time period was evidence of a social care need. As explained above, the predictor variables were derived from the first two years of data and were used to predict social care cost in the third year.

In developing the models, the first exploratory runs included all of the independent (predictor) variables, before selecting the subset of variables that seemed to be significant in order to develop more parsimonious models. In creating local models for each site we selected variables that gave the best local fit.

DEPENDENT (OUTCOME) VARIABLE

Predictive models used in the health sector typically aim to predict either future healthcare costs, or admissions or readmissions to hospital. In this study, our initial aim was to see if we could identify people who would start requiring intensive social care use, so our dependent ('outcome') variable needed to classify people according to whether or not they became users of intensive social care within the target year.

The initial definition of intensive social care was whether or not a person was admitted to a residential or nursing home or started receiving intensive levels of home support (defined as more than ten hours a week, or night-sitting). This group of users was therefore a subset of the broader group of people receiving intensive social care and did not include, for example, those people who received high levels of social care throughout the study period. From our original population of social care users, only a small minority fell into this subset of people (see Table 5.1): from a pool of over 300,000 social care records, less than 5,000 cases moved into intensive social care.

Table 5.1 Selecting change in social care status, Sites A to E

		Numbers
Social care utilisation records		342,627
Users receiving a social care service		78,522
Users aged over 75* receiving a social care service		46,782
Users aged over 75 receiving a social care service in 2007		30,586
Users aged over 75 receiving a high-intensity social care service in 2007		13,082
Users aged over 75 who change into high-intensity social care in 2007		4275
Users aged over 75 who change into high-intensity social care in 2007 or for whom expenditure on low-intensity social care increases by £5,000		5,361
Of which:	Site A	433
	Site B**	1,853
	Site C	1,913
	Site D	308
	Site E	854

* Excluding people with no age recorded on social care

** Note: not all of these are in the areas of Site B covered by the model

It became clear that this initial choice of dependent variable (i.e. a person who had an episode of intensive social care in the final year but not in earlier years) was, in practice, overly restrictive. We attempted simply modelling admissions to a residential or nursing home (leaving home care aside), but this did not prove much better. Finally, we opted to increase the number of cases by using a broader definition, namely either a first period of high-intensity social care in the target year (that is, none in previous years) *or* a significant increase in overall annualised costs above £5,000.

Most of the testing work was undertaken on this model using a £5,000 a year threshold. However, we also explored using the same criteria but a lower cost threshold of £3,000 or £1,000 a year. In this report we refer to these different models as the £5K (original), £3K or £1K models. Finally, as well as predicting commencement of intensive social care (i.e. social care 'events'), we also tested a series of models that predicted changes in social care costs.

SUMMARISING MODEL RESULTS

The accuracy of a predictive model can be described in terms of its sensitivity and specificity (see Box 5.1), and according to the positive predictive value (PPV) and negative predictive value (NPV). These are the same criteria on which the accuracy of any screening test is measured. The PPV is also known as the 'precision rate', and it is the proportion of patients identified by the model as being high risk who were correctly categorised.[20] In other words, PPV reflects the probability that a positive test reflects the outcome that the model is attempting to predict (care home admission, for example). Screening tests also use techniques such as the receiver operating characteristic curve (ROC curve) and the use of the C-statistic. The C statistic is the probability that the model will place a pair of people in the right order, giving a higher risk score to the person who is admitted to a care home than to the person who is not. The score varies between a maximum of C=1.0 for perfect prediction to C=0.5 for purely random predictions.

Box 5.1 Sensitivity and specificity[20]

Sensitivity

This is equal to the number of true positives divided by the sum of true positives and false negatives. A sensitivity of 100 per cent means that the test recognises all sick people as such. Thus in a high-sensitivity screening test, a negative result essentially rules out the disease. Sensitivity alone does not tell us how well the test predicts other classes (that is, about the negative cases). Sensitivity is not the same as the PPV (which is the ratio of true positives to combined true and false positives), the latter being as much about the proportion of actual positives in the population being tested as it is about the test.

Specificity

This is equal to the number of true negatives divided by the sum of the true negatives and the false positives.
A specificity of 100 per cent means that a screening test recognises all non-sick people as non-sick (i.e. healthy). Thus a positive result in a high-specificity test essentially confirms the disease. The specificity alone does not tell us how well the test recognises positive cases. We also need to know the sensitivity of the test to the class, or equivalently, the specificities to the other classes. A test with a high specificity has a low Type I error rate. Specificity is sometimes confused with the precision or the PPV, both of which refer to the fraction of returned positives that are true positives. The distinction is critical when the classes are different sizes. A test with very high specificity can have very low precision if there are far more true negatives than true positives, and vice versa.

The accuracy of a model that predicts a binary outcome (for example, admitted or not admitted) can be displayed in the form of a 2x2 table (see Table 5.2a).

Table 5.2a Describing the performance of a diagnostic of a model

		Predicted result		
		Yes	**No**	
Actual	**Yes**	True positive (TP)	False negative (FN) Type II error	Sensitivity =TP/(TP+FN)
	No	False positive (FP) Type 1 error	True negative (TN)	Specificity =TN/(TN+FP)
		PPV=TP/(TP+FP)	NPV=TN/(TN+FN)	

PPV = positive predictive value
NPV = negative predictive value

However, in order to compare multiple sets of results we have chosen to display the cells in a single line (Table 5.2b).

Table 5.2b Format used in this report*

	Predict 'No'		Predict 'Yes'		PPV	Sensitivity	Specificity
	Actual 'No'	Actual 'Yes'	Actual 'No'	Actual 'Yes'			
	TRUE NEGATIVE	FALSE NEGATIVE	FALSE POSITIVE	TRUE POSITIVE			
Model	TN	FN	FP	TP	TP/(TP+FP)	TP/(TP+FN)	TN/(TN+FP)

*Note: In Table 5.4 an additional column showing the C-statistic has been added

PREDICTING A MOVE INTO INTENSIVE SOCIAL CARE (£5K MODEL)

The basic models for each of the five sites were built using the variables drawn from the list outlined in Table 5.3. The subsets of variables in the final models were those where there was a significant relationship to the predicted outcome. In selecting variables we also aimed to maximise the number of predictive variables from the sites while retaining a degree of consistency across sites by using variables that were available from all five. So, for example, the base models did not include detailed GP information because this was only available from two of the five sites.

Table 5.3 Variables used in basic models

Type of information	Variables used
GP registration	Age and gender used to identify potential users
Previous inpatient activity admissions	No. emergency admissions No. emergency 'avoidable' admissions No. emergency 'medical' admissions No. emergency 'mental health' admissions No. emergency 'alcohol and drug' admissions No. emergency 'cancer' admissions No. admissions where the patient ultimately self-discharged No. ordinary elective admissions No. day case (elective) admissions Presence of chronic disease (27 variable) Diagnosis derived disease groupings (21 variables) Hierarchical condition categories (70 variables and 6 interaction terms) Average number of episodes per spell
Outpatient	Specialty of attendances (45 variables) No. different specialties
A&E records	No. A&E visits, arrival by ambulance No. A&E visits subsequent transfer to specialist No. A&E visits with X-ray investigation No. A&E visits with a medical (non injury) diagnosis
Social care	Needs variables: Visual hearing impairment Social isolation Access to informal carer Activities of daily living/functional markers Other health problems (reported by social care) Prior social care use: Respite care costs Low-intensity home care (<2 hours): number of days in year Medium-intensity home care (2–10 hours): number of days in the year Equipment and adaptations: number of episodes started in the year Day care: number of days in the year Meals: number of days in the year Individual recorded as having had an assessment during the year

Having created the linked datasets, logistic regression was used to generate the 'beta coefficients' for each of the input variables. These are estimates of the strength of the relationship for each variable, having standardised the variables so that they all have variances of 1. Finally, these coefficients were applied to the validation sample in order to measure the performance of the model according to its PPV and sensitivity. If the PPVs for the development and validation samples were similar (to within a few percentage points) then we were reassured that the model was relatively stable. In this report we have only presented the most predictive and most stable models for each site.

Table 5.4 summarises the results from the base models using a risk threshold of 0.5 (i.e. a risk score of 50). Two important indicators of the predictive accuracy of these models are their sensitivity and PPVs. Overall, we found that the models performed similarly across the five sites. The best model in terms of sensitivity was that for Site A, which had 17 per cent sensitivity. In terms of PPV, our best model was achieved in Site E, where the PPV was 56 per cent. The values for the C-statistic ranged from 0.81 to 0.91: a range that indicates good discrimination and compares favourably with other models.

Overall, the predictive accuracy of our models was moderate: PPVs of about 50 per cent are respectable, but the sensitivities of around five to 15 per cent are lower than we would have hoped. Typically, our models only detect about one in ten of the people that actually do start using intensive social care. We think the main

reason for the disappointing sensitivities of the models is that significant changes in social care use are relatively rare. The very high specificity of our models reflects the very high proportion of true negatives (that is, the overwhelming number of people aged 75 and above who do not move into intensive social care). This phenomenon means that measures such as the specificity, NPV and area under the ROC curve are all extremely good for our models, but this does not necessarily mean that the models are meaningful in practice.[21]

To avoid any doubt, these models were highly statistically significant. For example, had we simply tossed a coin for each person aged 75 and above living in Site A then this coin toss would perform as a 'model' with a sensitivity of 50 per cent and a PPV of about 1.2 per cent. So our models are about 30 times better than chance.

Also included in Table 5.4 are the results for a pooled model. The variables in the pooled model were based on those variables that were significantly predictive in the local area models with an added dummy variable to indicate the site itself. The magnitude of the coefficients on this dummy variable reflected how different each site's data were from the pooled data. Although the dummy variables did prove significant (see Table 5.7), models in which this variable was omitted performed almost as well. Given that the use of dummy variables might limit the ability to apply the model in other sites (where we would not know the appropriate value for the dummy's beta coefficient) we have tended to focus on the models without the dummy variables.

Table 5.4 Summary results of base model (using a threshold of 50 per cent)

	Predict 'No'		Predict 'Yes'		PPV	Sensitivity	Specificity	C-statistic
	Actual 'No'	Actual 'Yes'	Actual 'No'	Actual 'Yes'				
	TRUE NEGATIVE	FALSE NEGATIVE	FALSE POSITIVE	TRUE POSITIVE				
Site A	15,058	412	95	85	47.2%	17.1%	99.4%	0.914
Site B	8,845	199	26	21	44.7%	9.5%	99.7%	0.883
Site C	75,358	1,939	162	105	39.3%	5.1%	99.8%	0.854
Site D	-	-	-	-	-	-		-
Site E	52,940	537	52	67	56.3%	11.1%	99.9%	0.808
Pooled model	152,183	3,165	356	201	36.1%	6.0%	99.8%	0.854
Pooled model with site dummies	152,189	3,167	350	199	36.2%	5.9%	99.8%	0.861
Site D with pooled variables	22,903	247	30	9	23.1%	3.5%	99.9%	-

The models that performed best were those that were calibrated locally. Where the pooled variable set was used on individual sites' data, the results were not as good as those reported in the table above. The exception to this is the case of Site D, where it was not possible to build a stable model, in other words, we were able to generate a development model but when we subsequently tested it on the validation data, the results proved inconsistent. We think the reason this occurred is that the number of events we were predicting was

particularly low in Site D. However, we show the performance of Site D using the pooled model that we have developed and we later reintroduce data from this site.

For comparison, the Patients at Risk of Re-hospitalisation (PARR) tool used by the NHS in England for predicting emergency readmission to hospital had a reported PPV of 65.3 per cent and a sensitivity of 54.3 per cent with a threshold score of 0.5.[4] The Combined Model results are not presented in the same way (see discussion in Section 7).

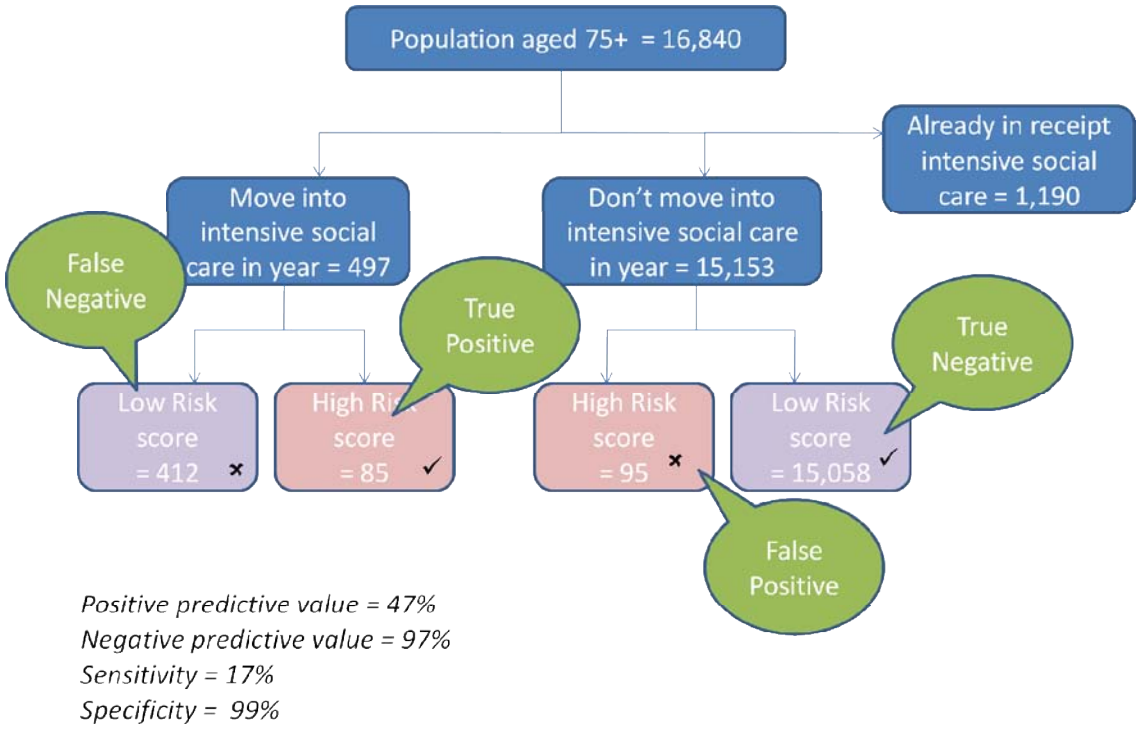
Summary observations:

We were able to generate models in four sites that predicted the start of intensive social care, as well as a pooled model that spanned all five sites.

Though the models have reasonable PPVs, most models have low sensitivity (in other words, they pick up only a small proportion of cases that move into intensive social care).

The biggest problem was that the numbers of events that the models tried to predict was small – it could be compared to looking for a needle in haystack.

Figure 5.1 Diagrammatic representation of how the model works in Site A



BEHAVIOUR OF INDEPENDENT VARIABLES

When performing multiple regressions, the strength of the relationship between each of the independent variable and the dependent variable is reflected in the beta coefficient. Positive coefficients denote that a variable is associated with an increase in 'risk'; negative coefficients with a decrease. A statistical test can be applied to determine whether or not the observed coefficients are simply due to chance. Therefore, by comparing the statistically significant coefficients it is possible to determine which variables have the largest influence on the model. It is worth noting again that for the majority of variables used in these models, the test statistics indicated a probability of less than one per cent (a standard test for statistical significance). Table 5.5 shows the beta coefficients of those variables that were significant at this $p<0.01$ level for all sites other than Site D.

Table 5.5 Variables in the basic models where the statistic was <0.01 (excluding age categories)

	Variable	Beta coefficient
Site A	Age band 7 (ages 85–89)	0.603
	Age band 8 (ages 90+)	1.207
	Social care data flag for health problem	2.094
	Any medium-intensity home care in past year	1.429
	Any day care in past year	1.792
	Count of A&E attendances, arriving by ambulance in past year	0.526
	No. different outpatient specialists in past year	0.171
	Count of A&E attendances, with medical diagnoses recorded	-0.54
	Constant term	-4.979
Site B	Sex = female	0.788
	Age band 7 (ages 85–89)	1.052
	Age band 8 (ages 90+)	1.429
	Any outpatient attendance in two years in specialty old age psychiatry	1.225
	Cost of respite care in past year	<0 .001
	Duration of medium-intensity home care in past year	0.003
	No. of social care assessments in last year	0.707
	Constant term	-5.64
Site C	Sex = female	0.327
	Age band 6 (ages 80–84)	0.826
	Age band 7 (ages 85–89)	1.124
	Age band 8 (ages 90+)	1.491
	Count of emergency admissions in prior 90 days	0.693
	Social care data flag for health problem	1.358
	Any medium-intensity home care in past year	0.724
	Any day care in past year	0.799
	Any social care assessments recorded in past year	1.203
	Any social care assessments recorded 24-12 months prior	0.463
	Constant term	-5.381
Site E	Any medium-intensity home care in past year	0.006
	Inpatient emergency admissions: ratio of no. episodes to no. spells in past year	0.367
	Any inpatient diagnosis in prior two years: rehabilitation	0.903
	Any meals supplied in past year	0.012
	Any day care in past year	0.016
	Any inpatient diagnosis in prior two years: urinary	0.695
	Count of A&E visits in prior 90 days	0.530
	Constant term	-5.06

In most cases, the coefficients are positive (indicating that a higher value on that variable is associated with a higher likelihood of the move into intensive social care). Though the significant predictor variables differ slightly from site to site, there are some common patterns, namely:

- Age is commonly an important factor (although oddly not in site E, apparently). In particular, ages 85 to 89 and 90 and above are important variables within the models.
- Where gender appears in the models, there is a positive association between being female and the risk of starting intensive social care.
- Markers of prior social care use are common and influential. These include medium-intensity home care (which features in all the models); day care; and the occurrence of social care assessments in earlier years.
- Previous emergency encounters with health services (either as admissions or A&E visits) also feature in most models.
- All the constant terms are negative and are around -5. This means that in the absence of any other variables, the beta terms sum to -5, and it results in a risk score that is very close to zero (0.007). The bigger the negative coefficient, the closer the risk score would be to zero. The constant term just reflects the fact that most people's risk scores will be nearer to zero than to 100.

The beta coefficients for the categorical age variables are shown for three sites in Table 5.6. These show the type of pattern that might be expected, with the higher age bands having an increasingly important weight in the model.

Table 5.6 Beta coefficients for age bands from three sites (relative to age band 75–79 = 0)

	Age band – beta coefficient value (significance)					
	80–84		**85–89**		**90+**	
Site A	0.31	(0.17)	0.6	(<0.01)	1.21	(<0.001)
Site B	0.68	(0.04)	1.05	(<0.01)	1.43	(<0.001)
Site C	0.83	(<0.001)	1.12	(<0.001)	1.49	(<0.001)

Table 5.7 shows the beta coefficients for the variables used in the pooled £5K models. All the age categories are significant and they increase in weight with increasing age. The dummy variables for the sites were also significant. In this case they are positive values relative to Site E. The set of variables describing social care use are clearly important drivers of the models, as is the 'other health flag' derived from the social care datasets. Of the health variables, the most important are:

- visit to A&E by ambulance
- emergency admission within the previous 90 days
- the average number of episodes per spell (an indicator of inpatient complexity)
- attendance at outpatient old age psychiatry clinics
- chronic disease flags for COPD, diabetes

Interestingly, many of the chronic disease markers are not significant in this model.

Table 5.7 Variables used in pooled models

Type of variable	Parameter	Estimate	Standard error	Probability
Constant	Intercept (constant term)	-5.8501	0.1022	<0.0001
Age and sex	Age band 6 (80–84) (relative to 75–79)	0.6105	0.0855	<0.0001
	Age band 7 (85–89) (relative to 75–79)	1.0073	0.0845	<0.0001
	Age band 8 (90+) (relative to 75–79)	1.2556	0.0888	<0.0001
	Sex = female	0.3546	0.0604	<0.0001
Site dummies	Site A dummy (relative to E)	0.7447	0.1034	<0.0001
	Site B dummy (relative to E)	0.4725	0.1312	0.0003
	Site C dummy (relative to E)	0.5111	0.0802	<0.0001
A&E data	Count of A&E attendances, arriving by ambulance in past year	0.2714	0.0426	<0.0001
	Count of A&E attendances, with medical diagnoses recorded	-0.1788	0.0696	0.0102
Inpatient data	Count of emergency admissions in prior 90 days	0.3778	0.066	<0.0001
	Any inpatient diagnosis in prior two years: angina	-0.1832	0.1365	0.1797
	Any inpatient diagnosis in prior two years: mental disorder	0.2582	0.1247	0.0385
	Any inpatient diagnosis in prior two years: Bipolar	0.0256	0.2488	0.918
	Any inpatient diagnosis in prior two years: Parkinson's	0.2089	0.266	0.4322
	Inpatient emergency admissions: ratio of no. episodes to no. spells in past year	0.1527	0.0309	<0.0001
	No of inpatient HCC flags in prior two years	0.0181	0.0338	0.5925
	Any inpatient diagnosis in prior two years: urinary	0.1707	0.1048	0.1035
	Any inpatient diagnosis in prior two years: diabetes	0.3581	0.1245	0.004
	Any inpatient diagnosis in prior two years: rehabilitation	0.014	0.1026	0.8915
	Any inpatient diagnosis in prior two years: COPD	0.4345	0.1346	0.0012
Outpatient data	No. different outpatient specialties in past year	-0.0213	0.0215	0.3198
	Any outpatient attendance in two years in specialty old age psychiatry	0.5363	0.1582	0.0007
Social care variables	Any social care assessments recorded 24-12 months prior	0.516	0.0841	<.0001
	Any social care assessments recorded in past year	0.9667	0.12	<.0001
	Any day care in past year	0.5925	0.107	<.0001
	Any low-intensity home care in past year	0.789	0.1075	<.0001
	Any medium-intensity home care recorded 24-12 months prior	-0.2772	0.1245	0.026
	Any medium-intensity home care in past year	1.6593	0.0863	<.0001
	No. of social care assessments in last year	-0.038	0.0668	0.5693
	Any meals supplied 24-12 months prior	0.2596	0.1505	0.0845
	Social care data flag for health problem	1.5464	0.1071	<.0001
	Respite care costs in prior year	0.000102	0.000054	0.0584

Summary observations:

The most important variables driving the prediction were those that described prior use of social care.

A number of health variables did appear to contribute to the models – but their effects were not large.

The number of significant variables in the final models was fairly small, making it possible to create reasonably parsimonious pooled models.

ALTERNATIVE MODELS: PREDICTING A CHANGE IN SOCIAL CARE USE

The definition of 'intensive social care' described earlier has the advantage that it focuses on the most expensive forms of care. However, it creates a problem in that the number of people moving into intensive social care in one year is relatively small. Our initial description of intensive care was based on:

- moving to a care home (nursing or residential)
- use of home care greater than ten hours per week
- annualised costs greater than £5,000.

Much of our work has been spent trying to improve the performance of these models in predicting these rare events. As an alternative, we changed the definition of the dependent variable so as to increase the number of cases available for detection. Specifically, we decreased the level of cost increase from £5,000 to either £3,000 or £1,000. It should be noted that this change also has wider implications when considering the potential utility of the models in practice.

Table 5.8 demonstrates how the number of cases we were aiming to identify decreased according to the magnitude of the change in costs we were aiming to predict. So in Site A, the original model was aiming to identify 497 people from a population of over 15,650 people aged 75 or over. But by reducing the costs thresholds to £1,000 this number increased to 849.

Table 5.8 Number of cases for different dependent variables

Site	People 75+	No. people with new significant costs or increase in costs of			Proportion of people 75+ with new significant costs or increase in costs of		
		£1,000	£3,000	£5,000	£1,000	£3,000	£5,000
Site A	15,650	849	604	497	5.4%	3.9%	3.2%
Site B	9,091	445	278	220	4.9%	3.1%	2.4%
Site C	77,564	3263	2396	2044	4.2%	3.1%	2.6%
Site D*	-	-	-	-	-	-	-
Site E	53,600	1194	818	605	2.2%	1.5%	1.1%

*Note: we were not able to develop a model for Site D that produced consistent results on the validation sample.

We then built a model using the pooled variable set (developed for predicting intensive social care at £5K) but now aiming to predict changes at the lower new thresholds of £3K or £1K. The results for the individual site models for £1K models are shown in Table 5.9. The best performing model was in Site E, with a PPV of 66 per cent and a sensitivity of 36 per cent. In Site D, there were still problems in building a reliable model.

Table 5.9 Results from the £1K model (pooled model variables)

	Predict 'No'		Predict 'Yes'		PPV	Sensitivity	Specificity
	Actual 'No'	Actual 'Yes'	Actual 'No'	Actual 'Yes'			
	TRUE NEGATIVE	FALSE NEGATIVE	FALSE POSITIVE	TRUE POSITIVE			
Site A	14,637	643	164	206	56%	24%	99%
Site B**	8,568	344	78	101	56%	23%	99.1%
Site C	73,885	2,772	416	491	54%	15%	99.4%
Site D**	22,503	544	84	58	41%	10%	99.6%
Site E	52,187	765	219	429	66%	36%	100%

** Note: the differences between the development and validation samples suggest over-fitting.

We used pooled data from four sites to predict a dependent variable based on a change of at least £1K a year or a move into intensive social care. Our independent variables were based on a pooled set of variables that were different from those of the individual area models (using the £5K threshold). Once again, these models exclude people who are already recipients of intensive social care.

Table 5.10 shows that using a threshold equivalent to a £1K a year produces a pooled model with a PPV of around 55 per cent and a sensitivity of 19 per cent. This model performs much better than the original (based on an increase of £5K a year). The model with a £3K threshold does not perform as well (it has a PPV of 42 per cent and sensitivity of ten per cent). Note for comparison that the pooled model described earlier had a PPV of around 35 per cent and a sensitivity of six per cent.

Table 5.10 Effects of changing the threshold on costs of social care

	Predict 'No'		Predict 'Yes'		PPV	Sensitivity	Specificity
	Actual No	Actual Yes	Actual No	Actual Yes			
	TRUE NEGATIVE	FALSE NEGATIVE	FALSE POSITIVE	TRUE POSITIVE			
Pooled model 5K	152,183	3,165	356	201	36%	6%	99.8%
Pooled model 3K	151,245	3,660	564	436	44%	11%	99.6%
Pooled model 1K	149,278	4,677	876	1,074	55%	19%	99.4%
Pooled model £1	143,598	8,154	1,559	2,594	62%	24%	98.9%
Pooled 1K, no SC cost Y2	144,056	2,824	100	57	36%	2%	99.9%

Looking at the individual coefficients from the £1K model (Table 5.11), they largely conform to what might be expected, namely:

- The coefficients for the age categories are all significant and positive, indicating increasing risk with increases in age.
- Prior social care assessment and utilisation are strong and significant predictors.
- Use of medium-level home care services in the previous year is associated with in an increased risk (although, interestingly, home care services from the year before that are negatively associated with risk).
- There are several significant healthcare usage flags: for example, the use of A&E (via ambulance) and emergency admission to hospital. The variable designed to indicate complex cases (ratio of episodes per spell, for example) was also significant.
- Some of the long-term condition flags are significant (for example, COPD, diabetes), while others were not.

Table 5.11 shows the important parameters in one variant of the pooled £1K model. As with the £5K model, the variables that were important predictors included those relating to prior social care assessment and medium- or low-intensity home care in the previous year. Of the health variables, prior A&E visits by ambulance and emergency admission in the previous 90 days were significant. This was a similar to our earlier models.

One important issue with all the models, but especially those with lower cost thresholds, is that a positive result (a transition to higher social care costs) may simply reflect a change in services towards the end of the year before the prediction year. So, for example, suppose a person starts to receive a new service in the last month of the year, their annual costs for that year may be below the £1K threshold, but if the intervention continues in the same way in the following year then their costs may exceed £1K. The model may rightly predict that their costs will change, but in effect the services they receive would be unchanged. We examined the extent to which this may occur in one of the sites by comparing the maximum monthly costs in the year before prediction with the prediction year itself. This suggests that this situation applies to around a quarter of those whose costs increase by £1K.

Table 5.11 Summary of beta coefficients (where p<0.05) for the pooled £1K model – with no dummy variable

	Variable	Beta coefficients	Probability
	Intercept	-4.96	<0.0001
Age and sex	Age band 6 (80–84) (relative to 75–79)	0.47	<0.0001
	Age band 7 (85–89) (relative to 75–79)	0.87	<0.0001
	Age band 8 (90+) (relative to 75–79)	1.00	<0.0001
	Sex = female	0.36	<0.0001
Social care prior use	Any social care assessments recorded 24-12 months prior	0.59	<0.0001
	Any social care assessments recorded in the past year	1.43	<0.0001
	Any day care in the period 24-12 months prior	1.09	<0.0001
	Any low-intensity home care in the past year	1.14	<0.0001
	Any medium-intensity home care recorded 24-12 months prior	-1.22	<0.0001
	Any medium-intensity home care in the past year	2.35	<0.0001
	Social care data flag for a health problem	2.14	<0.0001
	Any meals supplied 24-12 months prior	0.33	0.02
	No. of social care assessments in the last year	-0.14	0.03
Healthcare	Any A&E visit arriving by ambulance in the past year	0.25	<0.0001
	No. of emergency admissions in the past 90 days	0.29	<0.0001
	Outpatient visit in the past two years to the specialty of old age psychiatry	0.40	0.01
	Number of different outpatient specialties seen in the prior two years	0.07	<0.0001
	Ratio of inpatient episodes to admissions in the past year	0.16	<0.0001
	Any inpatient diagnosis of diabetes in the previous two years	0.39	<0.0001
	Any inpatient diagnosis of COPD in the previous two years	0.39	<0.0001

Summary observations:

Models that predict a smaller change in social care use are more successful, with better PPV and sensitivity.

Variables relating to prior social care use are again the most important predictors for these models.

These models do not identify many cases that were not previously known to social care.

CHANGING THE RISK THRESHOLD

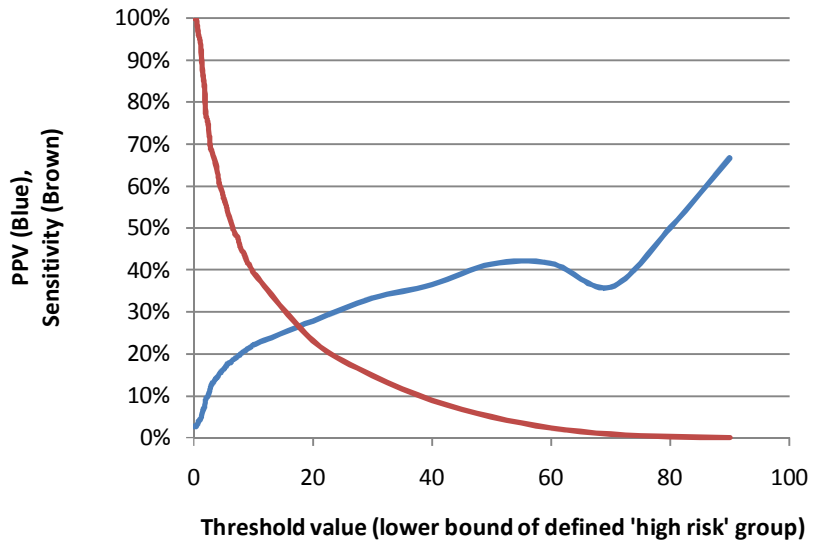
In the discussion above we have used a single cut-off point of 50. Where an individual's risk score is higher than 50 we have stated that the model was predicting a future change in social care use; and where it was less than 50 we have stated that the model was not predicting a future change in social care use for that person. While the choice of 50 is standard in these situations, it is possible to change this threshold in order to alter a model's performance.

We can plot the effect of changing the risk thresholds on PPV and sensitivity (see Figure 5.2a and Figure 5.2b). In general, using a lower threshold (say a risk score of 40) identifies slightly more cases (higher sensitivity) but with a corresponding increase in false positives (lower PPV and lower specificity).

However, the site-specific models produce highly skewed distributions of risk scores, with a large proportion of cases having a score of zero or thereabouts. This reflects the low sensitivity of these models. As a result, changing the threshold does not make much difference to the performance of these models.

In contrast, with the £5K pooled model, changing the cut-off to lower values results in a linear fall in PPV and an increase in sensitivity (more cases are being identified but the proportion of true positives falls). The maximum PPV of about 60 per cent is obtained at a threshold above 90. It is important to remember that there are very few cases in this range. So although making the definition of risk more stringent results in more accurate predictions, the number of positive cases identified (i.e. the sensitivity) falls even further.

Figure 5.2a Trade-off between PPV (**blue line**) and sensitivity (**brown line**) according to different risk cut-offs (pooled 4-site 5k model)

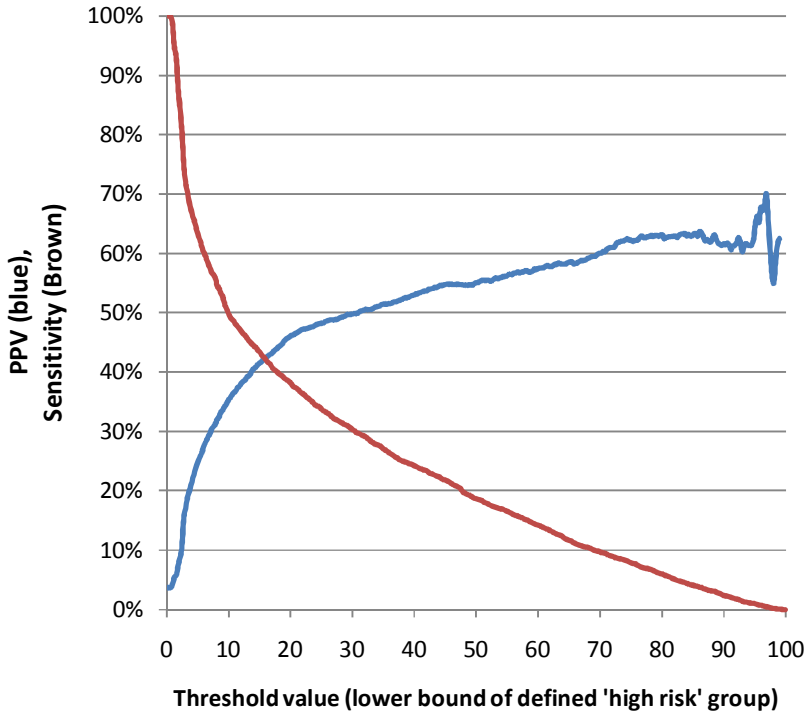


The equivalent plot for the pooled £1K model (Figure 5.2b) shows high values for both PPV and sensitivity across the risk spectrum. The sensitivity declines as the risk threshold is raised, but in this model the PPV drops away at using threshold values below 20. This means that using a threshold of around 20 would produce results with a sensitivity of 40 per cent (in other words, the model would identify four out of every ten people

who had an increase in social care costs) and a PPV of 40 per cent. With a risk threshold set at 70, the model would be more accurate (PPV of 60 per cent) but it would identify a smaller proportion of cases (sensitivity of about ten per cent).

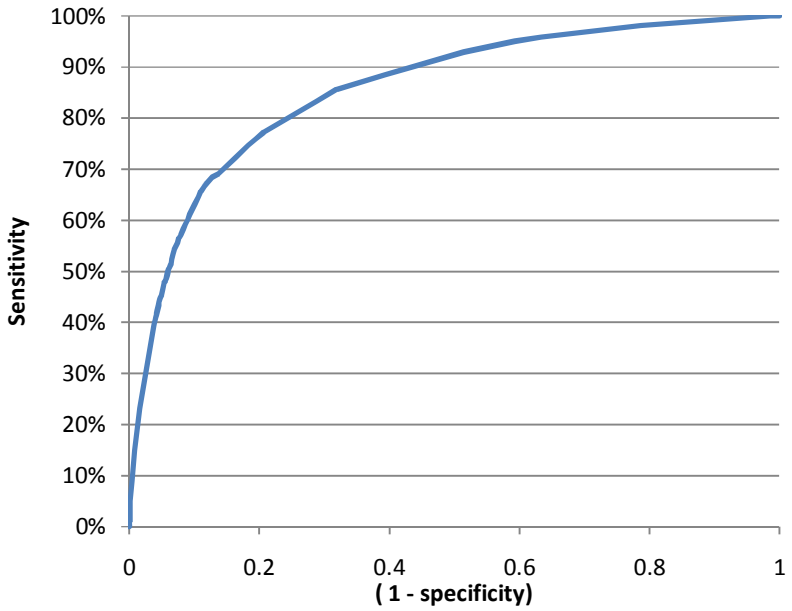
Figure 5.2b shows how these trade-offs might be used in practice when considering the costs and effectiveness of alternative interventions offered to people in different strata of predicted risk.

Figure 5.2b Trade-off between PPV (**blue line**) and sensitivity (**brown line**) according to different risk cut-offs (pooled 4-site 1k model)



An alternative graphical approach often used for presenting this type of information is the ROC curve. A ROC curve plots the trade-off between the sensitivity of the model and its specificity. The more cases detected (i.e. more sensitive), the more likely that there will also be a larger proportion of false positives. For a completely ineffective test, the relationship between these two parameters would be a straight line at 45 degrees to the axes. The more the curve raises above this diagonal line, the better the model. Figure 5.3 shows the ROC curve for the Site B model. The area between this curve and the 45° line can be calculated ('area under the ROC curve') and it is used as a measure of how well a model performs. For our models the areas under the ROC curve are generally quite high (equivalent to the C-statistic in the cases) – above 0.8 and indeed beyond 0.9 in one case (see Figure 5.4). Such values suggest that our models perform well. However, these figures might be somewhat misleading with models which attempt to predict events that are very rare in the population.[21] In this instance, there are so few cases with a positive outcome that the ability to predict when something doesn't happen is relatively easy – though the models do it well. For this reason, we have opted to concentrate on the PPV and sensitivity values.

Figure 5.3 Receiver operating characteristic curve (ROC curve) for Site B (C-statistic = 0.854)



Summary observations:

There is a trade-off between PPV and sensitivity. By changing the threshold value used to define high risk, it is possible to trade off the accuracy of the model with its sensitivity. A more sensitive model tends to be less accurate and vice versa.

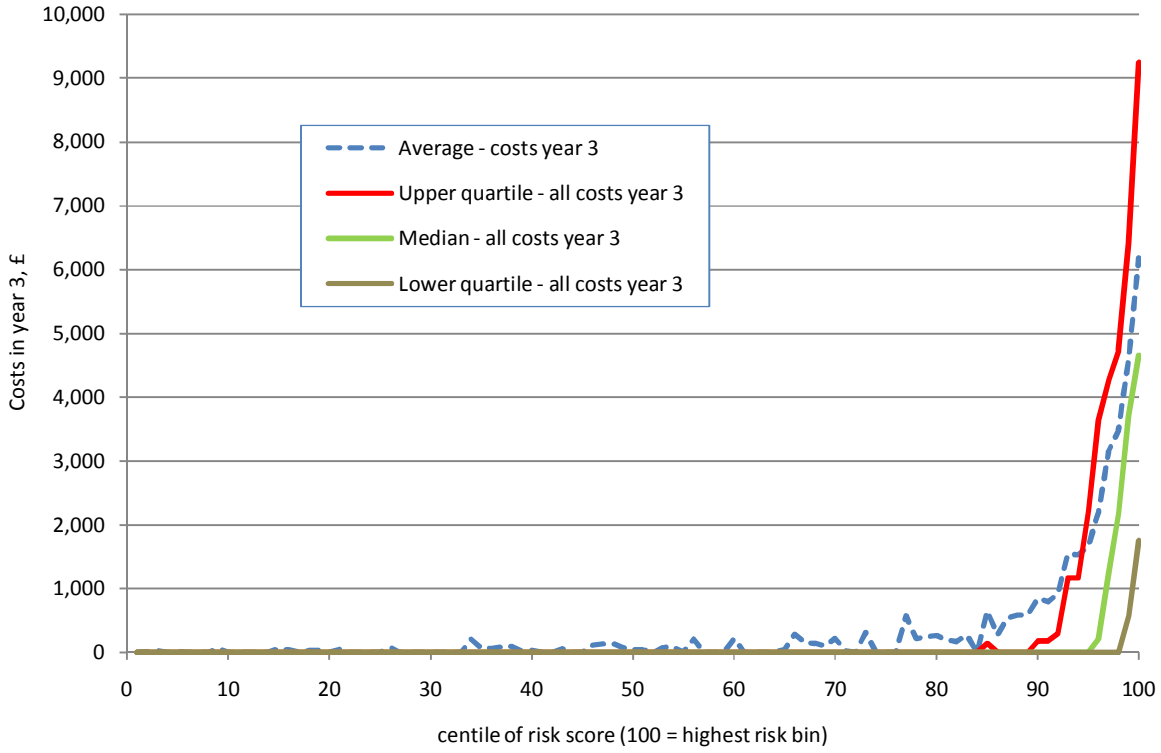
The models perform very well according to the parameters that are often used to describe screening tests, such as the area under the ROC curve.

PREDICTING COSTS – RISK SCORES AND ACTUAL COSTS

The pattern of annualised social care costs is highly skewed, such that only a relatively small number of cases with higher risk scores account for a significant majority of the costs.

Figure 5.4 shows the relationship between risk score (divided into 100 bins) and cost in the subsequent year. All of the models shown are highly skewed, and for about three quarters of cases there is no discernible risk score and accordingly no subsequent costs. So in practice we find that the costs are associated with only a small subset of our total population. The future costs increase in line with the risk score as we would hope, but the rise only really starts for the top seven to eight per cent of cases. The graph shows that the distribution of costs at any one point can be quite wide. On average, however, the models perform as we would expect, with increasing risk associated with increasing costs.

Figure 5.4 Site B £5K model: observed costs in Year 3 by centiles of predicted risk



6. VARIATIONS TO THE BASE MODELS

This section presents the results from a subset of the many different model combinations we developed in the course of this project. In some cases, the analyses are based on one or two of the five sites used for this study; in others they use pooled site models. Many of these experiments were attempts to improve the fit of our earliest models, rather than describing systematically the relationships between variables.

SOCIAL CARE ONLY

In Section 5 we noted the importance of social care information in the models. In light of this, we attempted to build a model that only used predictive variables from the social care data (as well as age and sex). We still used the registered GP population as the member file, but we removed any health variables from primary or secondary care.

Table 6.1 presents a summary of the performance of these models on a pooled dataset and compares to equivalent results from Section 5 (i.e. with the health variables included). It is clear in terms of case detection that these performed as well as the models that include health variables. This means that most of the discrimination of the model results from the social care variables. The health data only adds some marginal refinement to the predictions of the model, in the same way that social care data add predictive power to models predicting health outcomes (discussed later).

Table 6.1 Summary of model results using social care data only

	Predict 'No'		Predict 'Yes'				
	Actual 'No'	Actual 'Yes'	Actual 'No'	Actual 'Yes'	PPV	Sensitivity	Specificity
	TRUE NEGATIVE	FALSE NEGATIVE	FALSE POSITIVE	TRUE POSITIVE			
Pooled 1K model social care only	149,247	4,630	907	1,121	55.3%	19.5%	99.4%
c.f. pooled 1K model (with health variables)	*149,278*	*4,677*	*876*	*1,074*	*55.1%*	*18.7%*	*99.4%*
Pooled 5K model social care only	152,165	3,183	374	183	32.9%	5.4%	99.8%
c.f. 5K model pooled (with health variables)	*152,183*	*3,165*	*356*	*201*	*36.1%*	*6.0%*	*99.8%*

Table 6.2 shows the beta coefficients found for the £1K social care only model. Most of these variables were highly statistically significant and highly positive. So, for example, the model shows that the following are all positively associated with future increases in annualised costs: prior use of low-intensity home care, prior assessment and home care. The other health flag (a marker of health problems coded within social care records) is also significantly positive.

Table 6.2 Pooled £1K model using social care data only (no site dummies)

Variable	Beta estimate	Standard error	Probability
(Intercept)	-4.7103	0.0597	<0.0001
Age band 6 (80–84) (relative to 75–79)	0.4783	0.0655	<0.0001
Age band 7 (85–89) (relative to 75–79)	0.8738	0.0647	<0.0001
Age band 8 (90+) (relative to 75–79)	0.9397	0.0677	<0.0001
Sex = female	0.3073	0.0478	<0.0001
Any social care assessments recorded in the 24-12 months prior	0.6479	0.0692	<0.0001
Any social care assessments recorded in the past year	1.4417	0.0998	<0.0001
Any day care in the past year	1.1079	0.0958	<0.0001
Any low-intensity home care in the past year	1.2467	0.0949	<0.0001
Any medium-intensity home care in the 24-12 months prior	-1.3832	0.1194	<0.0001
Any medium-intensity home care in the past year	2.6201	0.0738	<0.0001
Social care data flag for a health problem	2.2184	0.0859	<0.0001
Number. of social care assessments in the last year	-0.0561	0.0611	0.3584
Any meals supplied in the 24-12 months prior	0.3014	0.1388	0.03
Cost of respite care in the past year	0.00002	0.000067	0.7339

ADDING DETAILED GP DATA

Two of the five sites were able to include data from GP clinical systems that record and code a wide range of health and social problems.

Three different variants of GP variables were tested. First, we tested a group of 33 high-level categories (see Table 6.3). For any person in our member file, we built 66 possible variables spanning any of these 33 groups for Years 1 and 2 of our models. However, given the limited number of cases we were trying to detect, it became necessary to reduce the number of variables to a minimum. We therefore replaced this very rich range of information with a small number of simple flags that indicated whether a person had had any three or more of these 33 variables in the previous two years.

Second, we reproduced 11 of the variables developed for the Combined Predictive Model.[22] These were developed for predicting emergency admissions to hospital, and so were potentially less appropriate for predicting social care admissions. In each of the two sites, only the three most significant Combined Model variables were entered into the models (see Table 6.5).

Table 6.3 Read Code groupings showing number of codes and 'events' mapped

	Site B		Site D		Both sites	
GROUPING	No. mapped Read Codes	Count of events	No. mapped Read Codes	Count of events	Count of events	% of all
No group	40082	26,191,161	56071	75,678,066	101,869,227	91.8%
Diabetes	190	280,303	231	1,406,544	1,686,847	1.52%
Hypertension	19	347,218	19	1,191,354	1,538,572	1.39%
Asthma/COPD	9	236,515	11	972,094	1,208,609	1.09%
Depression	17	285,926	17	575,126	861,052	0.78%
Heart failure or heart disease	5	164,538	5	522,070	686,608	0.62%
Heart disease/angina	9	99,972	9	321,581	421,553	0.38%
Anxiety (tranquilisers)	69	123,306	74	290,033	413,339	0.37%
Malnutrition	153	108,886	252	241,541	350,427	0.32%
Osteoporosis	7	118,191	9	189,284	307,475	0.28%
Psychosis	138	82,483	199	205,707	288,190	0.26%
Atrial fibrillation	3	71,646	3	212,128	283,774	0.26%
Anaemia	3	52,810	3	131,393	184,203	0.17%
Urinary incontinence	57	45,514	69	96,508	142,022	0.13%
Parkinson's disease	73	31,151	82	101,079	132,230	0.12%
Mental health	7	19,071	12	75,221	94,292	0.08%
Home visits	16	7,216	20	74,857	82,073	0.07%
Glaucoma	6	23,034	6	57,419	80,453	0.07%
Obesity	18	25,627	17	51,672	77,299	0.07%
Stroke	15	17,023	14	41,015	58,038	0.05%
Mobility	13	9,274	15	38,209	47,483	0.04%
COPD	15	19,116	12	16,651	35,767	0.03%
Autoimmune disease	1	15,021	1	14,937	29,958	0.03%
Falls	14	9,611	17	17,890	27,501	0.02%
Neurological disease	1	4,875	1	13,710	18,585	0.02%
Social care service indication	7	1,525	8	10,295	11,820	0.01%
Dementia	35	4,171	32	6,564	10,735	0.01%
Dependence (on carer/other)	4	1,797	4	7,237	9,034	0.01%
Isolation	3	1,180	3	4,518	5,698	0.01%
Nursing home/other	5	1,648	5	3,069	4,717	0.00%
Bowel Incontinence	5	820	8	1,468	2,288	0.00%
Confusion	6	1,451	4	828	2,279	0.00%
Dehydration	2	131	2	272	403	0.00%
Blindness/deficiencies of vision	1	85	1	243	328	0.00%

Table 6.4 Combined model variables used in modelling for each site

Site B		Site D	
Combined model variable name	**Definition**	**Combined model variable name**	**Definition**
discnt7pl_flg	Seven or more distinct disorders in prior three years	**gp_drug36_365:**	Bronchodilator preparation in prior 180–365 days
gp_poly_0509_123	Polypharmacy: 5–9 unique drugs in prior three months	**gp_dis48:**	Psychotic disorder in prior two years
gp_poly_10pl_123	Polypharmacy: Ten or more unique drugs in prior three months	**gp_dis47:**	Psychoactive substance misuse disorder in prior two years

Table 6.5 shows the effect of each of these groups of variables on model results. The overall findings were that the GP data made a marginal difference to the performance of the models – the Combined Model variables performed least well, but the sensitivity and PPV remained fairly stable.

Table 6.5 Effect of adding GP data to models

	Predict 'No'		Predict 'Yes'				
	Actual 'No'	Actual 'Yes'	Actual 'No'	Actual 'Yes'			
	TRUE NEGATIVE	FALSE NEGATIVE	FALSE POSITIVE	TRUE POSITIVE	PPV (%)	Sensitivity (%)	Specificity (%)
Site B £1K model plus 3 combined model variables	8,574	345	72	100	58	22	96
Site B £1K model plus 2 GP flag counts	8,572	338	74	107	59	24	96
c.f. Site B £1K model with no GP variables	***8,573***	***341***	***73***	***104***	***59***	***23***	***96***
Site D £1K model plus 3 combined model variables	22,539	561	48	41	46.1	6.8	99.8
Site D £1K model plus 2 GP flag counts	22536	557	51	45	46.9	7.5	99.8
c.f. Site D £1K model with no GP variables	***22,538***	***556***	***49***	***46***	***48.4***	***7.6***	***99.8***

In addition, we used GP data when applying the Adjusted Clinical Groups™ (ACG) system to our models. In undertaking this work, we were supported Johns Hopkins University who kindly allowed us to try their proprietary software package and offered us advice and support.

ACGs™ are themselves part of a family of different classification systems and it is important to stress that in this particular example we tested only the additional value of the ACGs™ in the particular situation we were studying. ACGs™ have a much wider application than in predicting a subset of changes in social care. Like a wide range of additional variables (including deprivation, GP data and community services data), ACGs™ did not add significant predictive power to our models.

SOCIOECONOMIC VARIABLES

We would expect an individual's uptake of council-funded social care to be related to their own individual socioeconomic circumstances. For example, the funding of social care by local authorities is means- tested so this may well have an important impact upon the patterns of use seen both within and between populations.

In this study we were not able to measure individual socioeconomic status directly. Instead we had to use the proxy characteristics of the area where the person lived. The use of area-based analyses is common in predictive modelling, and it can be relatively sophisticated, particularly if the areas studied are small. However, in order to preserve confidentiality, the datasets we received for this study only had relatively broad area descriptors.

In our first runs of the models we used the Index of Multiple Deprivation (IMD). Published by the Department for Communities and Local Government, this serves as a proxy metric to describe a person's socioeconomic status. The IMD index is applied at the Lower Super Output Area (LSOA) level but we only received LSOA codes of residence from two of our five sites. These were Site D (which we did not include in our pooled model due to poor performance) and Site B (our smallest site). So, instead we decided to use the four-site pooled model to test IMD, where we were obliged to use the IMD score for the GP practice instead of LSOA. The assumption here is that an individual person is represented by the average IMD for people registered at the practice. In fact, we were only able to model using IMD on three sites because in the missing site we had difficulties mapping many members of the registered population to a unique GP practice.

We tested models using the IMD score of the GP practice as a continuous variable, as a categorical variable (high, medium, low) and finally as a simple flag to indicate an area of high deprivation. In the latter case the beta coefficient was statistically significant; however, the overall performance of the model was not drastically improved.

Next we tested the behaviour of a model restricted to people from areas of high deprivation. We did this to explore if means testing for social care had a notable impact on our predictive accuracy. Our hypothesis was that in areas of high deprivation, the effect of means testing would be less important so that a model that was unaffected by means testing would perform better than elsewhere. Again, we found that this approach did not radically improve our models.

Table 6.6 Effect of adding deprivation scores

	Predict 'No'		Predict 'Yes'				
	Actual 'No'	*Actual 'Yes'*	Actual 'No'	*Actual 'Yes'*	PPV (%)	Sensitivity (%)	Specificity (%)
	TRUE NEGATIVE	FALSE NEGATIVE	FALSE POSITIVE	TRUE POSITIVE			
c.f. 3 site model without IMD	*99,307*	*2,623*	*237*	*138*	*36.8*	*5.0*	*99.8*
3 site model with flag for highest quartile IMD score	99,293	2,622	251	139	35.6	5.0	99.7
Model restricted areas with highest quartile IMD score	25,124	777	102	55	35.0	6.6	99.6

We agreed to test some alternative information on social and economic characteristics at the small area level by using the Mosaic™ datasets created by Experian. The Mosaic™ variables we were able to use only represent a small subset of the possible data. However, we were obliged to keep the number of additional variables to a minimum because of the small number of cases being predicted.

Table 6.7 shows the results of the modelling with these Mosaic™ variables included. As can be seen, the addition of these variables did improve the models and in fact the performance diminished slightly. This should not necessarily be seen as reflecting the relationships between these variables and transitions to social care. Rather, the way we used the information was necessarily crude because it was limited by the small numbers of cases we had and by the need to map information to relatively large geographic areas. We suspect that these types of variables will probably add more predictive power to models for areas larger than a single primary care trust/local authority site.

Table 6.7 Results of £1K model including MOSAIC™ variables, Site D

	Predict 'No'		Predict 'Yes'		PPV	Sensitivity	Specificity
	Actual 'No'	Actual 'Yes'	Actual 'No'	Actual 'Yes'			
	TRUE NEGATIVE	FALSE NEGATIVE	FALSE POSITIVE	TRUE POSITIVE			
C.f. (Site D 1K best)	22,538	556	49	46	48.4%	7.6%	99.8%
+ Mosaic™ variables	22,539	562	48	40	45.5%	6.6%	99.8%

COMMUNITY HEALTH SERVICE DATA

Additional information on healthcare use is potentially available from community services (such as district nursing). Community healthcare is an important factor for many people receiving social care and the data may identify lower levels of interventions than major acute hospital episodes.

One site (Site D) was able to provide us with a large dataset that included 1.3 million community health services contacts for 55,000 patients over a period of five and half years. When linked to the social care data, it appeared that the majority of social care users did not receive community services in this period. Table 6.8 shows that only 27 per cent of the people receiving social care also received community healthcare.

The key variables from community healthcare data related to prior use of community health services, as well as any diagnostic information recorded by community services staff. Three variables were tested in the model (Table 6.8). These were selected as being the most promising in terms of frequency and the relationship with receiving more costly social care in Year 3. However, as can be seen from Table 6.8, the effects on the model results were disappointing in that performance dropped slightly.

Table 6.8 Community health service variables tested: relative prevalence in people aged 55+ for those going on to receive significant social care or a £1K increase in social care costs in Year 3

Variable	Of group with	
	Predict 'Yes'	Predict 'No
People with at least one district nurse contact in last year	27%	4%
People with at least one 'other' (not: DN, Spec. Nurse, Comm. Matron) contact in last year	39%	12%
People with diagnosis concerning mobility	18%	5%

Table 6.9 Impact of community services data when added to the predictive model

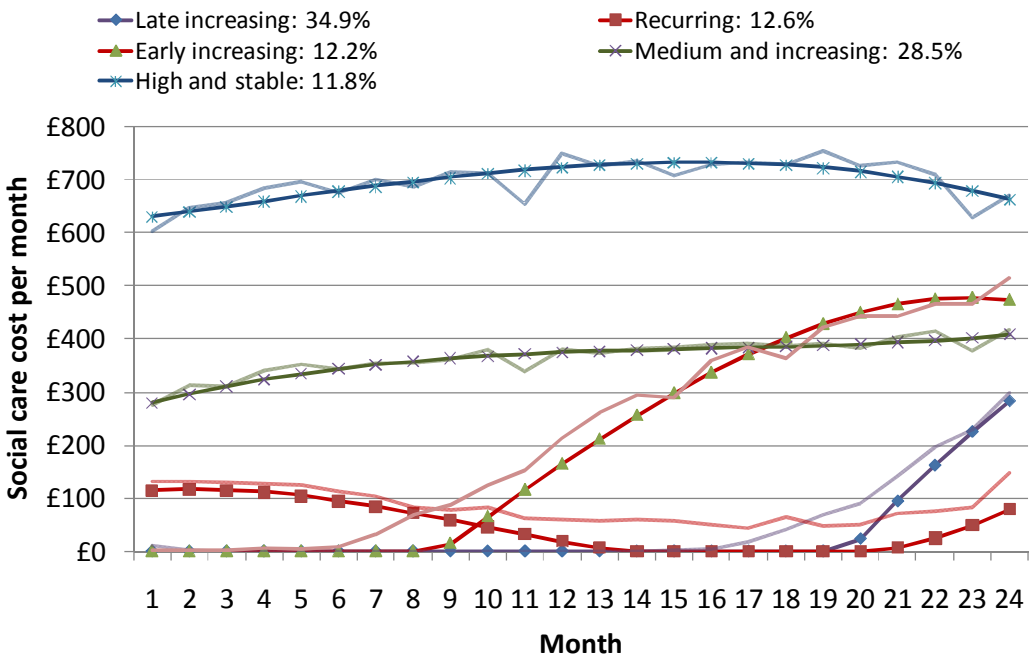
	Predict 'No'		Predict 'Yes'				
	Actual 'No'	Actual 'Yes'	Actual 'No'	Actual 'Yes'	PPV	Sensitivity	Specificity
	TRUE NEGATIVE	FALSE NEGATIVE	FALSE POSITIVE	TRUE POSITIVE			
(C.f. Site D £1K best)	22,538	556	49	46	48.4%	7.6%	99.8%
+ Community care variables	22,534	557	53	45	45.9%	7.5%	99.8%

EXPLORATORY ANALYSES OF TRAJECTORIES AND COSTS

One of the important strengths of the dataset we collated for this project was its ability to describe a sequence of health and social care for individual people. We believed that past contact with care services could provide some useful information about potential future use, which is why we tested and used such variables in our basic models. However, these variables were rather crude and often simply flagged whether a service was used in a particular time period. In order to exploit the longitudinal nature of the data more effectively, we examined the pathways (or 'trajectories') experienced by individuals.

Figure 6.1 shows the trajectory groups for those people who started high-intensity social care in Year 3, or who had an increase in social care costs of at least £5,000. It shows the observed mean social care cost for each trajectory group for each of the 24 months in the Years 1 and 2, together with the expected mean social care cost estimated from the fitted models. The observed and expected mean costs were similar in most cases, although there were some differences in the groups with the lowest costs. As can be seen from the figure, we assigned the following descriptive names to the five trajectories: 'Late increasing', 'Early increasing', 'High and stable', 'Recurring' and 'Medium and increasing'.

Figure 6.1 Trajectories of social care costs in Years 1 and 2 for people who start high-intensity social care in Year 3 (or have £5K or more increase in costs)



Our earlier models attempted to predict an event such as a move into intensive social care. The specification of these models was optimised for the purpose of case finding, which is similar to the Patients at Risk of Re-hospitalisation (PARR) tool and the Combined Model used in the health sector. However, another option might be to construct models that predict costs alone, either current or future. We undertook some exploratory work to examine the potential of cost prediction models (in contrast to the event-based models described earlier). These cost models represent a different technique that offers further potential though in different settings. We decided to design a model that aimed to predict *any* social care cost for those people who had received no intensive social care in the two previous years. We did this in two stages:

1. Our first step was to see if we could predict whether a person had any social care costs in Year 3
2. The second stage was to estimate the magnitude of those costs.

The models for the first stage used information on prior social care use and so, unsurprisingly, they were very good at predicting future use. For example, from Site D the model fit was very good, with an r-squared value of 0.4. Unfortunately, these statistics are a little flattering because they are largely driven by variables relating to prior social care use and so they would be of limited value in practice. These models are equivalent to case finding approaches that try to predict any use/expenditure in Year 3. The key drivers were prior use of social care and prior social care assessments.

Summary observations:

We tested many different variants of the basic model using a range of additional datasets.

Models predicting social care use based only on social care data performed almost as well as those including health data. Social care only models would be much easier to implement in practice.

The additional impact of detailed GP level information was assessed using some alternative classification schemes. None of these increased the performance of our models significantly.

Adding proxy markers of socioeconomic circumstances at an area level did not improve model performance appreciably, nor did the addition of community health service data.

Housing data and council tax and information proved difficult to obtain.

7. THE IMPLICATIONS FOR CASE FINDING

This section discusses the results from Sections 5 and 6 in terms of their implications for using predictive models as case finding tools within social care settings. This was the original intention of the feasibility study. We begin by discussing the model results and then go on to consider some important aspects relating to implementation.

OVERALL FINDINGS

In summary, the work on model development has found that:

- These models tried to predict rare events (i.e. changes in social care status within the target year). This made it difficult to build predictive models. Although most of the models we developed produced statistically significant results, there are still important questions of how these would add value to everyday practice.

- Those models that predicted smaller changes in social care use (the '£1K' models) searched for a larger number of cases and as a result they performed better than the original ('£5K') models, which sought to detect much steeper, and much rarer, changes in resource use in a year (equivalent to the commencement of intensive social care such as admission to a care home).

- Most of these models performed passably well in terms of their positive predictive value (PPV). While the PPVs were lower than those for the Patients at Risk of Re-hospitalisation (PARR) tool (which predicts unplanned readmissions to hospital), they did compare favourably with some published findings on the combined predictive model (which predicts unplanned hospital admissions).

- The £5K models lacked sensitivity, in that they only identified a small proportion of cases from the population at large. In contrast, the £1K models, which predicted smaller changes in social care costs, performed better in terms of sensitivity. In fact, the lower the cost threshold, the better the model.

- The models produced broadly similar results in each of the five sites. Models constructed with local data performed better than pooled models constructed with data from more than one site. It was, nevertheless, possible to create a useful model based on pooled data from four of the sites.

- The most important predictor variables were those based on information regarding prior social care use and social care needs. Models built from social care data alone performed roughly as well as those that contained health and social care data. Nevertheless, certain health variables were significantly predictive of future social care costs.

- A wide range of variants and refinements were tested models but none of them produced dramatic improvement in the performance of the models. Variants included:
    - use of the Index of Multiple Deprivation (IMD) as a predictor variable
    - constructing models based only on residents living in 'deprived' areas
    - addition of GP data (in two sites)
    - use of community healthcare data (for example, district nursing, community physiotherapy)
    - classifying people according to Adjusted Clinical Groups™ (ACG) and related groupings.

There are a number of factors that may have reduced the accuracy of predictive models. They fall broadly into three categories:

1. problems with the data
2. problems with our methods
3. problems with reality (i.e. the real world that our models are trying to represent).

PROBLEMS WITH THE DATA

The review of the literature identified a number of variables known to be associated with the use of social care by older people that we then sought to acquire from routine data. In some instances, however, the data were not recorded in a sufficiently consistent manner to be usable. One problem was that some of the most important predictive variables are, by their very nature, the hardest to capture in structured data systems. For example, the loss of an informal carer, or the fact that somebody's family has moved out of an area, may be the critical factor leading to a move to residential care. Such events are unlikely to be stored in electronic form except, perhaps, within free text.

We attempted to obtain additional information about people's social situations from other local authority datasets but had limited success and were unable to obtain council tax data. We therefore recognise that our models are not capturing some of the most important predictive variables. This would be consistent with the observation that our best models are reasonably specific but are insensitive (the predictions they make are accurate but they only detect a minority of cases from the population at large).

Given that all the information was extracted from operational information systems, our analyses had to exploit data that were less than perfect. All electronic systems of this type are susceptible to slight variations between individuals ('noise') in terms of the completeness and consistency with which information is recorded. Moreover, for some variables, we have had to undertake fairly complex manipulation in order to construct standardised variables. It is possible that we have inappropriately interpreted the meaning or the values of certain data fields. Once again, these inconsistencies make it more difficult to create a robust model. In fact, at the outset there was a view that these challenges might be so great that we would be unable to derive any kind of useful model. Any work to improve the consistency and recording of important information, especially in social care, would be of benefit for this type of work.

PROBLEMS WITH THE CHOICE OF ANALYTICAL TECHNIQUES

The analytical approach was based on our earlier work in the field of predictive modelling. Previous models, such as PARR, have been built in situations where the number of events being predicted was markedly higher, and where the datasets were derived from more homogeneous sources. Given more time and more data, we believe there are many other analytical avenues that we could pursue in the prediction of social care use.

The use of a split-sample methodology whereby one dataset is used to develop the model and another to validate it – is a fairly common technique. This step is necessary to avoid over-fitting the model – when the relationships between variables are unique to the datasets studied and so the model is not reproducible on other data. An alternative would have been to use a computationally more intensive 'bootstrapping' process in

which repeated samples of data are used to construct many different models – and the consistency of model weights are then assessed.[23, 24]

Alternatively, we could have looked again at the choice of time scales which was fairly crude. In our models we sought to predict events in the next 12 months. However, we would be interested in exploring whether changing this time frame produces more accurate predictions. In particular, we note that our models generate large numbers of false negatives (people who appear to have little previous health and social care information but who suddenly require intensive social care). So it may be that if we concentrated on a time span that was closer to the event of interest, say three months rather than 12 months, then we might be able to build more accurate models. We believe that a more thorough analysis of the pathways of care (illustrated in Figure 8.1 in the following section) may help us to assess the significance of the best time frames to use.

The choice of attempting to predict a change in social care status created a challenge all of its own. In particular, we needed to categorise people across systems and across services using imperfect information. This meant that we had to classify users in ways that aimed to make an observed change in one site roughly equivalent in scale, context and causality to a similar change in the other sites. To do this, we were required to make assumptions that the information recorded about service levels was comparable across sites; that the consistency and recording of data were the same; and that the triggers leading to changes in service were comparable. An alternative approach might be to construct models that were more closely tailored to local data. It is interesting to note that at least one commercial company advocates recalibrating its predictive models for hospital admissions according to the characteristics of local datasets.

PROBLEMS WITH CONTEXT

Aside from the technicalities of the datasets and the methods used, there are important questions about whether the thousands of local decisions that underpin routes into care are sufficiently consistent to make predictive modelling feasible. Any predictive model for social care will have to capture the consequences of individual professional decisions, the local context, and wider policy decisions at council level. In general, predictive models do not require uniformity. However, we suspect that the degree of variability in decision-making in this field – where so many choices are made by local councils and individuals – may be too wide to allow common patterns to be discerned.

This project aimed to predict episodes of high-intensity social care funded by the local authority. We did not receive data on self-funded episodes of care (those not funded by the local authority), so this was not available to us for modelling. Some individuals will self-fund their care for a period of time but later require financial support from the local authority (typically because they have 'spent down' their assets and now begin to qualify for council-funded care under the means-testing rules). Our predictive models may not be identifying such individuals if the start of their high-intensity social care funded by the local authority was not preceded by observable patterns in social and healthcare utilisation (40 per cent of people newly admitted into high-intensity social care in Year 3 received no social care funded by the local authority in Year 2). It is not possible for us to say which of these individuals had been previously self-paying for care. However, it is noteworthy that in some sites we were able to predict admission to local authority-funded support for some of them, based on healthcare variables.

Equally, it is possible that some of the false positive cases generated by our models do in fact go on to receive high-intensity social care, but pay for it without funding from the local authority. Unfortunately, we did not

have data on the incomes and assets that might have been used to predict whether or not someone might become eligible for means-tested support.

MOVING FROM THEORY TO PRACTICE

The initial aim of this project was to explore the feasibility of developing predictive models for social care that were analogous to the case finding tools used by the NHS for predicting emergency hospital admissions and readmissions (such as the Combined Model and PARR, respectively). Although there are many other potential applications for predictive models (discussed below), the principal focus of our efforts was on developing a case finding tool for social care.

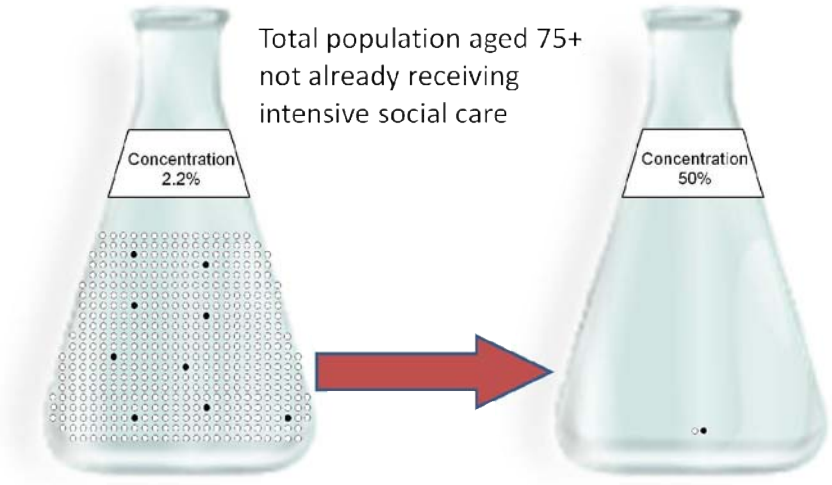
If these models were to be used in practice, we would need to consider a number of practical considerations. These include factors such as information governance, business planning, concurrent use with PARR/Combined Model, as well as operational issues such as running the model locally or centrally.

ADVANTAGES AND DISADVANTAGES OF DIFFERENT MODELS

When we use a threshold of 50, our models have reasonably good PPVs but a disappointingly low sensitivity. Their effect can be considered rather like a distillation process in which we move from a population where the events of interest are very rare, to a distilled population where they are much more concentrated.

For example, of the people aged 75 and above not already receiving intensive social care, only 2.2 per cent of people will require intensive social care next year. In contrast, of the individuals flagged by the predictive model, a full 50 per cent will require intensive social care. In other words, those people identified by the model are 23 times as likely to require intensive social care as the average person aged 75 and over who is not already receiving intensive social care (see Figure 7.1).

Figure 7.1 Distillation of high-risk individuals



It is also important to note that sensitivity can be traded off against PPV by altering the threshold of the models after they have been built, and that the accuracy of our models compares favourably with other predictive models currently being used in the United Kingdom (see Tables 7.1a and 7.1b).

Table 7.1a Comparative accuracy of our models with those currently used in the NHS (expressed as risk thresholds)

Model	Risk threshold	PPV (%)	Sensitivity (%)
PARR (England)	50	65.3	54.3
	70	77.4	17.8
	80	84.3	8.1
SPARRA (Scotland)	50	59.4	18.0
	70	76.1	3.3
Pooled £5K	50	41	5
	70	36	1
Pooled £1K	50	55	19
	70	60	10

Table 7.1b Comparative accuracy of our models with those currently used in the NHS (expressed as % of ranked population)

Model	Ranked population	PPV (%)	Sensitivity (%)
Combined model (England)	Top 1%	40.5	6.0
	Top 20%	15.9	47.1
PRISM (Wales)	Top 1%	44.3	6.6
	Top 20%	16.5	48.8
Pooled 5k	Top 1%	35	13
	Top 20%	10	75
Pooled 1k	Top 1%	57	15
	Top 20%	14	76

Another point worth emphasising is that the cost of the event being predicted in social care may be considerably higher than the cost of a hospital admission. For example, a typical emergency hospital admission might cost £2–3,000, whereas the average cost of a year of social care will often be £30–40,000. In other words, the stakes may be a lot higher for social care than for healthcare. This means that a viable business case might be constructed for early intervention in social care even if the accuracy of the social care models is not as high as for the health models (and even if the cost of the intervention was higher and/or its effectiveness lower than for preventive health interventions).

Despite high-quality evidence that some interventions can prevent or delay care home admissions,[25, 26] it is less clear how cost-effective these interventions would be when applied in practice for the people identified by the model as being at high risk. For example, it may transpire that some of the people identified by the models have immitigable risks. As a starting point, we would need to compare the characteristics of such people with the inclusion criteria for the interventions in question. The relatively high PPVs of our models are attractive to commissioners because relatively little preventive money would be spent unnecessarily on false positives. Indeed, some of the 'false positives' might actually experience the event the following year. The problem with the low sensitivity of our models is that we are unable to reach most of the people who experience the adverse event, so the impact on the population as a whole would be severely limited. Although it may still make business sense to invest in preventing a small number of these costly adverse events, it does raise some ethical concerns in terms of the justice of spending large amounts of resources only on people identified by the model.

One hypothesis for the low sensitivity of our models is that many people who require intensive social care appear 'out of the blue' when they have spent down their savings to the point of requiring financial support from the local authority. Such people may have spent many months or years in receipt of social care, perhaps being kept so well that they have little contact with the council or the NHS, and therefore have few predictor variables for us to analyse. So the addition of information about people who self-fund their care would be potentially very helpful.

One decision to be made by a local authority is over the trade-off between PPV and sensitivity discussed above. During our discussions with the sites, we found that some local authorities were interested in implementing very low-cost interventions (posting brochures and information leaflets, for example). These sites would probably want to set the threshold at a low level in order to maximise the sensitivity of the model. Other sites said that they envisaged using the model as a case finding tool for their multi-disciplinary team of social workers, physiotherapists and occupational therapist. For such an intensive, costly intervention it would be important to maximise the PPV at the expense of sensitivity.

INFORMATION GOVERNANCE

At the outset of this project, we obtained permission to analyse pseudonymous data and to report back in aggregate form, but not on individual cases. We liaised with the Patient Information Advisory Group (PIAG) who advised us that if the data we used were pseudonymous (i.e. effectively anonymous to us because we did not have access to the key to decrypt the scrambled unique key) then this project could be conducted without recourse to Section 60 of the Health and Social Care Act 2001. From PIAG's perspective, the nub of the issue was whether or not we were able to derive the identities of people recorded in the data.

However, having developed working models, some of the councils are very enthusiastic about starting to use them and have asked us whether we can give them back the pseudonymous identifiers of the individuals that

have been flagged up as being at risk. The sites would like to re-identify the individuals (i.e. convert the pseudonyms back into real names and addresses) in order to contact these high-risk people, identify their needs, and perhaps design and offer them preventative care.

The use of case finding models in the health sector was facilitated by guidance that PIAG published in November 2006 that set out clear requirements for primary care trusts (PCTs) or strategic health authorities wishing to run the PARR/Combined Model.[27] It included advice on the need to publicise how the population's data would be used; on effective pseudonymisation; and on ensuring that the first point of contact with patients was made through a clinical team already known to the patient (for example, the patient's GP practice). For the PARR/Combined Model, PIAG has advised that the initial contact with high-risk patients can either be made through a letter from the practice (stating that the GP wants to refer the patient to a new service and asking them to make an appointment with the GP surgery to discuss it), or by suggesting the referral during the course of a routine consultation.

We believe it will be important for PIAG's successor, the Ethics and Confidentiality Committee of the National Information Governance Board of Connecting for Health (ECC) to issue analogous guidance for the new social care models that we have developed. Any such guidance will have to address the additional complexities of data sharing and linkage across health and social care, over which there is currently still a degree of uncertainty.

CHOOSING THE RIGHT MODELS

As part of this project we tested a whole suite of different models. There are a number of factors to consider when operationalising these results.

Health versus social care models: It is likely that many of the sites that wish to use our new social care models will also be those who have the capacity and inclination to run the Combined Model. As we have demonstrated, there is a degree of overlap between the two models. In addition, each model predicts a number of 'by-product' events, i.e. the social care model will predict some healthcare outcomes such as A&E visits in people at high risk of social care and vice versa. This is a reflection of the complexity of the highest-risk patients in terms of their health and social care needs.

By running the Combined Model and the social care model in parallel, it becomes possible to calculate the number of 'overlap' patients and the number of 'by-product' events. Using a presentations software package (Xcelsius software – see Figure 8.2) we have illustrated how this information can be used to construct a 'dashboard' for a local health and social care economy. This can be used to obtain a system-wide perspective of these overlap individuals and 'by-product' events and potentially to allow a PCT and local authority to invest jointly, pro rata to the expected savings.

Local versus pooled or national models: Another decision for policy-makers is whether to opt for a national ('pooled') model for social care, or to develop a series of local models. There are several potential advantages of using a pooled model, including:

- Economies of scale from building a single model rather than reproducing the same work in each site across England.
- Improved generalisability (since the model would be built on a very large sample).
- Ability to run the model through a secure website akin to the PRISM model that operates in Wales[28] (which should increase the likelihood that the model would be run in practice).
- A pooled model would allow the risk profiles of sites to be compared across the country.

The main disadvantage of a pooled model is that its predictive power will typically be lower than that of a model built solely on local data. Given that social care data are recorded idiosyncratically across the country, the loss of predictive power may be even greater than in healthcare, since the social care predictor variables have to be collapsed considerably for inclusion in a pooled model. A local model would be able to exploit all of the available variables from social care and from other local data sources (housing and community healthcare data, for example).

Choosing the basic datasets: Our initial aim for this project was to create models that drew on information held in both health and social care information systems. This was based on the expectation that both data sources would contribute predictive power regarding which people were most likely to move into intensive social care. We wanted to use as comprehensive a set of variables as practicable and so we included GP and community healthcare data where these were available.

In retrospect, it has become clear that the effort involved in obtaining, linking and analysing some datasets cannot be justified by the impact they have on the final predictions. Although our best models all included some health variables, it is clear that much of the explanatory power could be obtained with social care datasets alone. In other words, it seems quite possible to use prior social care as the only data on which to create a risk score. Using this sole data source would undoubtedly be more straightforward in terms of the logistical and governance issues, but it would also undermine the additional benefits of using linked health and social care datasets.

In terms of healthcare data, we were unable to derive additional predictive power from the GP clinical data or from the community healthcare data. This might be a reflection of the way in which we handled these particular datasets but our current conclusion is that a simple GP registration file (Exeter file) and the hospital datasets are the only healthcare datasets that are needed for case finding in social care. Note, however, that there may be other reasons for incorporating GP data and community healthcare data, such as for performing gap analyses, for displaying all contacts with the health and social care services, and for assessing the impact of preventive interventions on all parts of the local health social care economy.

Recalibration: Our experience of the PARR tool suggests that analysts, managers and clinicians in local sites all require support in understanding, running and refreshing the predictive tool. For the PARR tool, the King's Fund held a support contract with the Department of Health. Queries were received via the King's Fund website and these were either dealt with directly or passed to Health Dialog UK if the nature of the query was more technical. We believe that a similar helpline would be needed if predictive models for social care were to be introduced.

SOFTWARE DEVELOPMENT AND SUPPORT

Another lesson from the PARR/Combined Model project is that, in order to avoid confusion and misconceptions, we would recommend paying close attention to the branding of the predictive models. The names chosen for the PARR model caused some confusion because:

- PARR was subdivided into PARR1 and PARR2 (depending on whether all hospital readmissions were to be considered or only those relating to an ambulatory care sensitive condition) but the Combined Model was sometimes erroneously referred to as PARR3 because it was in fact Phase 3 of the overall project. However, the Combined Model was quite distinct from PARR1 and PARR2 because it predicted hospital admissions rather than readmissions.
- Adding to the confusion were the different software versions for PARR, namely PARR, PARR+ and PARR++

Overall, however, the term PARR has widespread brand recognition within the NHS – more so than the Combined Model, which is a generic term also used by other model developers.

In terms of the software with which the models are implemented, there are various options that could be used, each with its own advantages and disadvantages (see Table 7.2). On balance, we would probably recommend option 3, although option 4 is a clear ambition for the future since it would ensure that predictions were 'pushed out' to front-line practitioners.

Table 7.2 Software options

	Software interface	Example	Advantages	Disadvantages
1.	None	Combined model	• Low cost centrally	• Relies on local skills • Duplication of effort • 'Cottage industry' approach to emailing predictions to sites in Excel spreadsheets
2.	Dedicated software	PARR	• Moderate cost centrally • Increased uptake (for example, PARR much more popular than the Combined Model partly because of free software)	• High cost centrally • Ongoing cost of upkeep of the software and software support
3.	Secure website	PRISM	• Moderate cost centrally • Changes rolled out rapidly • Available on clinician's desktop	• Clinician still has to 'pull' the result from the website
4.	Integrated within the electronic medical record	None	• Prediction is 'pushed' to the clinician	• Has not been developed in practice in the UK • Might need several versions for each of the main suppliers of GP EMR • Would not reach hospital clinicians • May also need a version for managers/commissioners in PCT

8. OTHER APPLICATIONS OF LINKED HEALTH AND SOCIAL CARE DATA

This section explores some of the other potential uses of linked data of the type we have developed in this project. Although our brief was to test the feasibility of developing a case finding model for social care, we believe that the potential for other applications is large and that it merits further exploration. Indeed, some of these applications may actually be of greater strategic importance than predictive modelling for case finding.

PRESENTING LINKED HEALTH AND SOCIAL CARE INFORMATION

As part of this project we have demonstrated the ability to collate data relating to service use for individual people over time. This includes all contacts with primary care, secondary care, community healthcare and social care. Displaying this information graphically provides an extremely rich picture of a person's interaction with the entire health and social care economy.

Presenting this 'raw' information to professionals in graphical form would require explicit consent from individuals, but could be very powerful. It could help professionals build a complete picture of a person's pattern of service use over time. Given that computer systems within the NHS do not usually talk to each other, let alone between the NHS and social care, clinicians must currently rely on the patient's recollection of healthcare use ('past medical history') and social care use ('social history'). These graphical representations could serve as a useful prompt and might help ensure a much more complete history.

As with predictive risk scores, ideally this descriptive information would be presented to clinicians within the computer systems that they use day-to-day, rather than requiring them to download from a specific program or website. This is because we suspect that the information is far more likely to be used if it is 'pushed' to the professionals rather than requiring them to go to the effort of 'pulling' it specifically. The information we have collated will probably be most useful to GPs, social workers, A&E doctors and hospital doctors (especially in acute medicine, general medicine and medicine for older people). When people come into contact with healthcare and social care services it can often be in a haphazard way, resulting simultaneously in both a duplication of some aspects of care and in incompleteness in others. Presenting information in this new graphical way may help generate a more seamless experience for clients.

Likewise, if we are to move towards a system where health and social care costs and benefits are potentially interchangeable, then the ability to characterise patterns of service use in this way will become increasingly important.

CONFIDENTIALITY CONSTRAINTS

Even where there are no personal details attached to the diagrams charting a person's contacts with health and social care services (no names, dates of birth or addresses), it might still theoretically be feasible to deduce the identity of the person simply by recognising a pattern of service use. For this reason, in presenting these diagrams in this report, we have introduced an element of random error to remove the possibility of

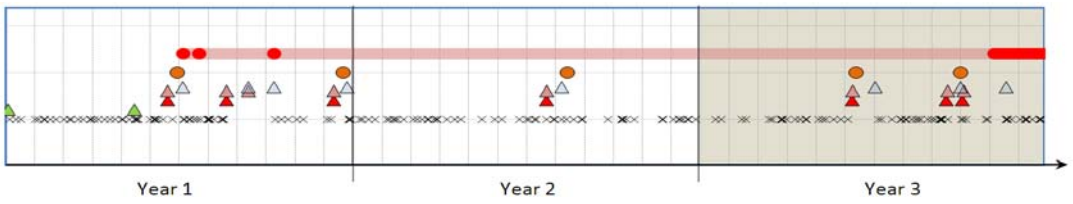
attributing details to an individual. When used in practice, however, we would need to make absolutely sure that these diagrams were only made available to clinicians directly involved in the care of the person concerned and with the explicit consent of the person.

Figure 8.1 Graphical representation of a person's interaction with the healthcare and social care services

PERSON A

Figure 8.1a shows an individual's health and social care history over a period of three years. In the first half of the first year there are many GP visits and a couple of attendances in outpatient clinics. Midway through that year, the person has an A&E attendance which results in admission to hospital. This hospital stay lasts a couple of weeks, during which time a social care assessment takes place. As a result of the assessment the individual is provided with a low-intensity package of home care. There are two further unplanned admissions that year (via A&E), with another three in the following two years. Social care assess the individual during or shortly after each emergency stay. After the final assessment shown, there is a further emergency admission, after which time a higher-intensity package of home care is put into place. Our model predicts a high-intensity event in this final year (shaded region) and this is indeed observed.

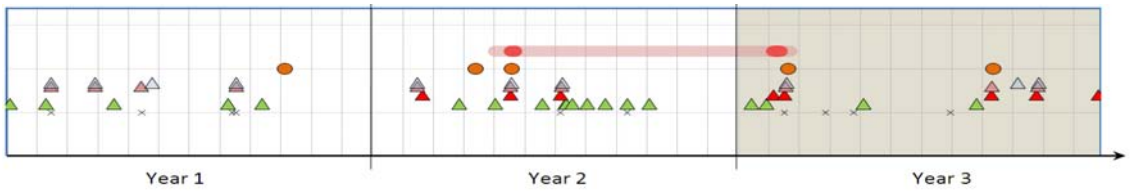
Figure 8.1a Health and social care event timeline for a 'true positive' individual



PERSON B

Figure 8.1b shows an individual's health and social care history over a period of three years. In the first year there are five outpatient attendances and four hospital admissions (as well as some GP visits). A social care assessment is carried out towards the end of the year, but this appears not to result in any service being provided. With more hospital contacts in the following year, two assessments take place and a low-intensity package of home care is put into place. Going into the third year (shaded region) our model predicts an increase in the intensity of the home care package or a care home admission. However, this is not observed: in the third year several unplanned hospital admissions occur, as well as a couple of social care assessments, but no social care services continue.

Figure 8.1b Health and social care event timeline for a 'false positive' individual.



INFORMATION FOR SERVICE USERS AND CARERS

Most of our discussion has been concerned with developing information tools for professionals. However, we suspect that services users (and potential service users) might also benefit from viewing shared information. For example, the ability to undertake comparisons with an expected pattern of resources, or a prediction based on a risk score might be seen as a valuable input to a person's own choices about questions relating to their care. Such questions might include:

- What sort of care do other people have who are in a similar position to mine?
- What is the likelihood that I will need more intensive care in the coming year?
- How much is my care likely to cost?

GAP ANALYSIS

A care gap is defined as a difference between the care received by a person and the recommended optimal care. An example of a gap is a patient who has had a myocardial infarction but is not taking an anti-platelet medication such as low-dose aspirin. The number and nature of care gaps reflects an opportunity to improve a patient's care, in such a way that patients with many high-impact gaps could be prioritised for intervention. Clearly, there are an extremely large number of potential quality gaps. However, it may not be possible to deduce many gaps from routinely collected data alone (for example, did a patient who was admitted to hospital with COPD have their inhaler technique checked within 24 hours of admission?).

Gaps may be defined and weighted either using non-statistical techniques (expert opinion, for example), or by using evidence-based standards such as those published by Milliman.[29] Where evidence-based gaps are used, it may be possible to quantify the expected impact of closing each gap by using the published adverse event rates from studies of patients who did and did not have that gap.[30] This information could then be used to help construct more reliable business cases for preventive care.

In general, predictive models rely on positive signals in administrative data to make forecasts of healthcare use, for example, a new diagnosis or a visit to an emergency department. In contrast, gaps typically represent negative observations, such as no follow-up visit or the absence of a particular drug. Pulling together data from multiple sources helps to identify more potential gaps. The Quality and Outcomes Framework (QOF) identifies a number of potential gaps (for example, patients with diabetes not taking a statin), but QOF gaps only become apparent if the diagnosis in question has been recorded in the GP record. Linking data from hospitals means that if the patient had a diagnosis made in secondary care but this diagnosis was not transferred into the GP record, then the gap of not taking a statin would still be apparent.

When it comes to social care gaps, the evidence base is less comprehensive but there may still be certain standards agreed by expert opinion (for example, having a social care assessment annually if in receipt of intensive home care, or having an assessment by an occupational therapist whenever admission to a care home is being considered). One specific social care 'gap' that might usefully be identified from linked datasets is that of a lack of reablement. Reablement may be defined as,

> '... giving people over the age of 18 years the opportunity and confidence to relearn/regain some of the skills they may have lost as a consequence of poor health, disability/impairment or going into hospital or residential care, and to gain new skills that help them to maintain their independence'.[31]

People who experienced a 'reablement gap' could be identified by trawling routine data for those people who should have received reablement and then dismissing those people who did actually receive reablement (see Table 8.1).

Table 8.1 Logic for determining 'reablement gaps' from routine data

Should have received reablement	Did receive reablement	Experienced a reablement gap
A	**B**	**A minus B**
1. Aged ≥18 2. Have had a hospital admission[†] 3. Were then assessed as being eligible for a home care service[‡]	(a) Met the criteria in **A** (b) Received assistance (limited to 5–6 weeks) that developed their skills for more independent living. The assistance may involve a range of clinical, therapeutic and social interventions, which are more appropriately met outside the home environment[§] (c) Received regular reviews of their progress[§] (d) If the person has substantial ongoing needs that can be met by qualified medical and nursing staff, the Reablement Service will aim to complement these activities[§]	= **A minus B**

KEY GP member file
[†] Hospital Episode Statistics data
[‡] Social care data
[§] Social care data and/or community services data

BUSINESS PLANNING AND COST MODELLING

In the *Predicting Costly Care* report, a commissioning tool courtesy of Health Dialog UK was discussed. This illustrated how primary care trusts (PCTs) can build business models for preventive care interventions according to the predictions of the NHS combined predictive model. To use the tool, commissioners begin by selecting a segment of the population according to their predicted risk of unplanned hospital admission (top five per cent of risk, for example) and choose how much money they wish to invest in preventive care for these individuals (for example, £100 each). The panel at the bottom of the tool shows the predicted frequencies of four adverse outcomes that the preventive intervention is designed to avoid, namely: inpatient emergency admissions, inpatient non-emergency admissions, A&E attendances and outpatient visits. By estimating the effect of a proposed preventive intervention on each outcome (for example, expected 20 per cent drop in emergency inpatient admissions as a result of the intervention) the tool uses the known predictive accuracy of the Combined Model to calculate the savings that the intervention would yield in the coming year.

We have now developed a prototype commissioning tool for health and social care. Figure 8.2 gives one example. The tool uses software that can be embedded within Word, PowerPoint or Adobe Acrobat Reader, which allows commissioners to select patients from the combined predictive model (CPM), the social care predictive model (SCPM) or both. When patients are selected from both models, the model first identifies any 'overlap' patients who were identified by both models. This feature is important because it ensures that any savings are not 'double counted'. For each patient, the model calculates the number and cost of all the

adverse events below, regardless of whether the individual patient was identified from the CPM or the SCPM. This reflects the fact that although the Combined Model is designed to identify patients who experience a costly NHS event (emergency hospital admission), it will as a 'by-product' identify people who experience other costly NHS events as well as costly social care events. The same is true for the SCPM.

By listing a number of costly social care outcomes and healthcare outcomes for each person selected from one or both predictive models, it becomes possible to calculate the potential relative savings that the council and the PCT might enjoy given the effectiveness of a particular intervention in mitigating the risks of each outcome. We believe that this might encourage joint investment and the development of pooled budgets because any savings could be divided pro rata by the two organisations (noting how these relative proportions vary across different segments of risk). At present, our business planning tool does not take account of reduced social care costs for people who are in hospital, but this could readily be envisaged for a future version.

As well as predicting costly events (admission to a care home or the instigation of intensive home care), we have also modelled predicted social care costs *per se*. Knowledge about expected costs is useful for ensuring that resources are allocated efficiently, and it is a prerequisite for individualised budgets. Cost models can be used retrospectively to understand the degree of variance in expenditure, i.e. the extent to which an observed pattern of spending is predictable given the characteristics of the user population. A typical approach is to group patients according to common characteristics, for example, by classifying patients into diagnostic related groups that are based on the standard disease and surgical procedure classifications. Unfortunately, there is no analogous social care classification system in common use. We believe it might be possible to develop such a grouping based on our modelling approach, where homogeneous groups of people are bundled together according to their expected social care resource utilisation characteristics.

TOOLS FOR EVALUATION PERFORMANCE MANAGEMENT AND REGULATION

Comparative assessments of performance, such as those undertaken by the Care Quality Commission (the regulator of health and social care), need ways of ensuring that their analyses compare like with like. It is clear that, at an individual user level, there are significant variations in the patterns of both health and social care use according to the characteristics for the client. To take an obvious example, all our models showed significant increases in future use of intensive special care related to age but a number of diagnostic groupings were also associated with high-risk individuals. So, for fair comparisons to be made, we need ways to standardise these effects on case type.

The Department of Health is currently investing in a number of 'upstream' interventions aimed at averting 'downstream' costs. Examples include the Partnerships for Older People Pilots, the Integrated Care Pilots and the Whole System Demonstrators. When evaluating such interventions, it is critically important to adjust for the types of people seen by the intervention so as to ensure that any observed savings are real and not simply an artefact of (a) shifting focus, or (b) regression to the mean (see Box 8.1).

Figure 8.2 Illustration of a business planning tool-linking health and social care cost and activity[vi]

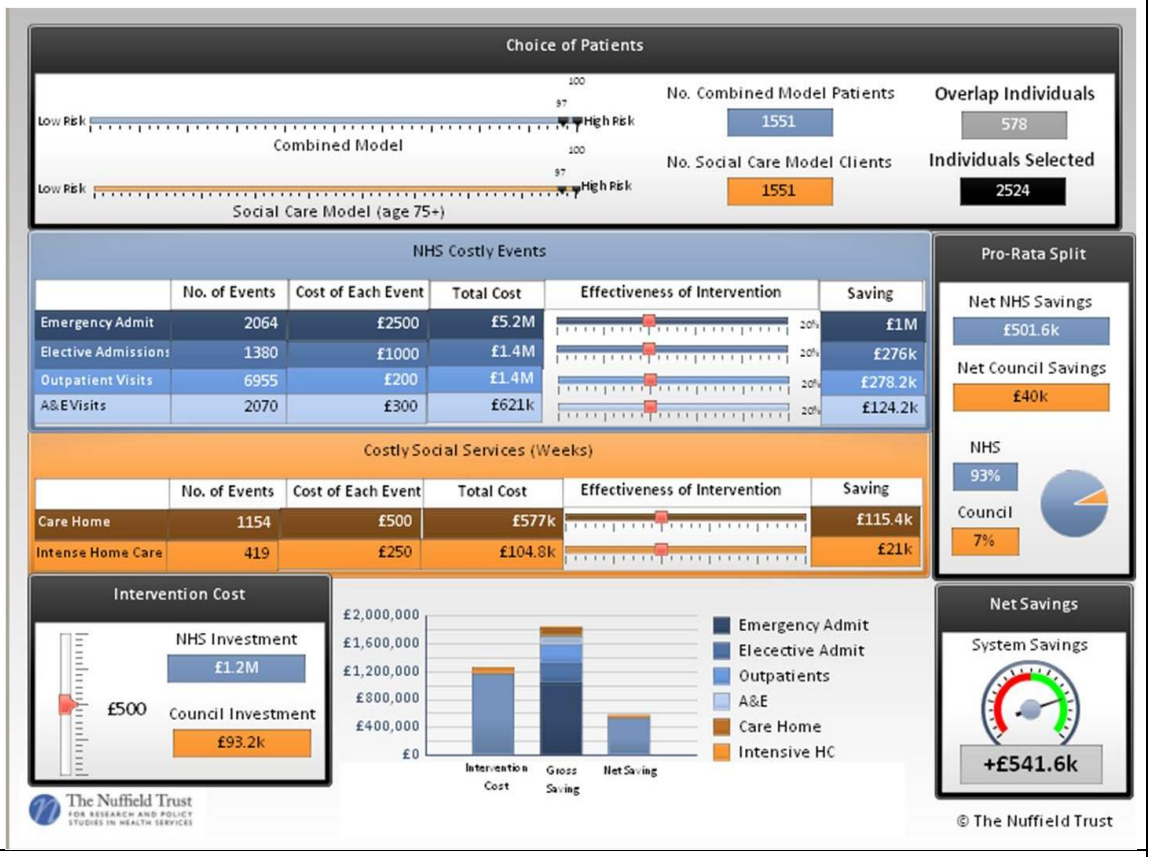


Figure 8.2 illustrates the situation when both the Combined Model and the social care model are being used to identify patients at the same time. The horizontal bars on Figure 8.2 are set to select the three per cent of the population at highest risk of social care costs according to the 5K model (1,551 patients), and the three per cent of the population at highest risk of emergency hospital admission being selected from the combined predictive model.

We can see that three per cent of the population will be 1,551 people from each model. Note that the dashboard signals that there are 578 'overlap' individuals. These are people who have been identified from both models. To avoid the problem of double-counting, the dashboard only counts these 'overlap' people once, so that the number of 'unique individuals' (2,524) is the sum of the rest of the number of people identified by both models minus the 'overlap' people (1,551 + 1,551 – 578 = 2,524).

The rest of the dashboard in Figure 8.2 relates to the 2,524 'unique individuals' identified as being at risk by either model but without double-counting. It is important to note that when using this dashboard, the two models can be used to select any risk strata of risk from either or both models. So rather than choosing the top three per cent of risk from both models, for example, we could have selected the top three per cent of risk from one model and the top six per cent of risk from another. Or we might have selected the 92nd to the 94th centiles of risk on one model and the 80th to the 85th centiles of risk on the other model.

With the current settings in Figure 8.2, we can see that the net savings to the NHS and the local authority are 96 per cent versus four per cent, respectively. This yields net savings of £476.2K to the NHS, £17.9K to the council and a total net saving of £494.2K for the local health and social care economy.

[vi] For demonstration purposes only, the sliders in this diagram can be moved to model the effectiveness of the intervention. (This function may not work on older versions of Adobe Acrobat).

Box 8.1 Potential artefacts in the evaluation of preventive services

Shifting focus	It is important to ensure that any reductions seen in hospital or social care utilisation are not simply due to a service shifting its focus towards lower-risk patients over time. For example, if a site had a declining number of higher-risk patients (perhaps because of the effectiveness of an intervention), then it is possible that the volume of services offered to lower-risk patients could increase. This would mean that, without risk-adjustment, the evaluation would underestimate the impact of the intervention on service utilisation because the impact on higher-risk patients would be obscured by the increase in services offered to lower-risk patients. This difficulty can be overcome by stratifying according to predicted risk, so that the impact on higher-risk and lower-risk patients will be assessed separately.
Regression to the mean	Evaluations must make sure that any apparent reductions in utilisation are not simply a statistical artefact caused by selecting high-risk patients for treatment. By selecting high-risk patients, there is a natural tendency for subsequent measurements on those patients to show reductions in use ('regression to the mean'). This means that if patients are chosen for an intervention based on their current high rates of hospital admissions, then we would expect their rates of hospital admissions to reduce over time *even in the absence an intervention*. Accordingly, without risk-adjustment, the evaluation would overestimate the effectiveness of the intervention on hospital use because some or all of the reductions observed would have happened anyway. This difficulty can be overcome by matching the people who received the intervention to people in other areas with the same level of predicted risk.

Similar techniques could be applied to the evaluation of reablement services, ensuring that any observed reductions in future hospital or social care use were true effects and that would not have happened anyway. So, for example, if we wanted to look at the experiences of older people following an emergency hospital admission, it would be possible for us to look at their subsequent use of social care. Figure 8.3 illustrates the average length of time between hospital discharge and admission to high-intensity social care, stratified for different age groups. A year after discharge from hospital, about 20 per cent of people aged 85 and over will have received intensive social care. This type of analysis shows:

- The potential for reablement to prevent people from moving into intensive social care after discharge from hospital.
- Reablement needs to be highly targeted because even in the highest age group, less than 20 per cent of people will move into intensive social care.
- Most people starting intensive social care do so within the three months of discharge from hospital – hence the need to act quickly.

Figure 8.3 Analysis of a cohort of older people showing the proportion not receiving high-intensity social care after discharge from hospital following an emergency hospital admission (stratified according to age).

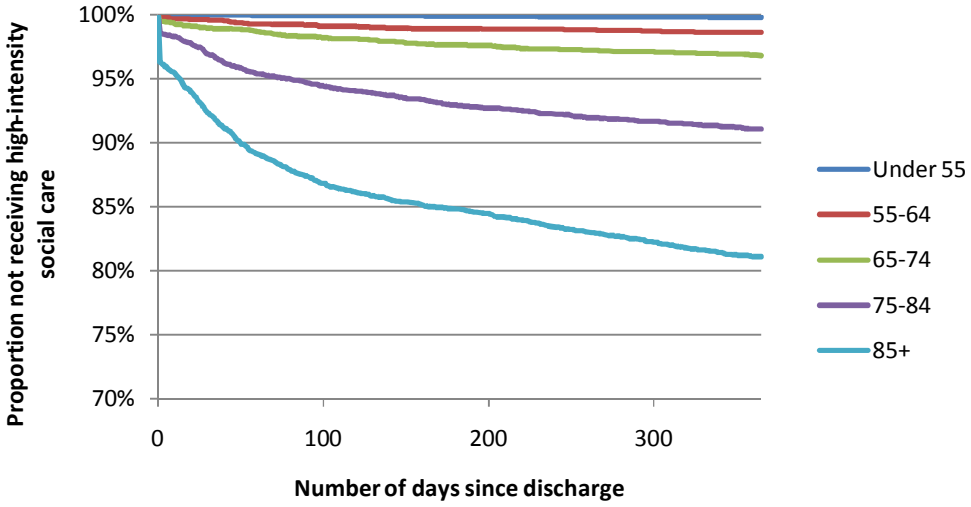
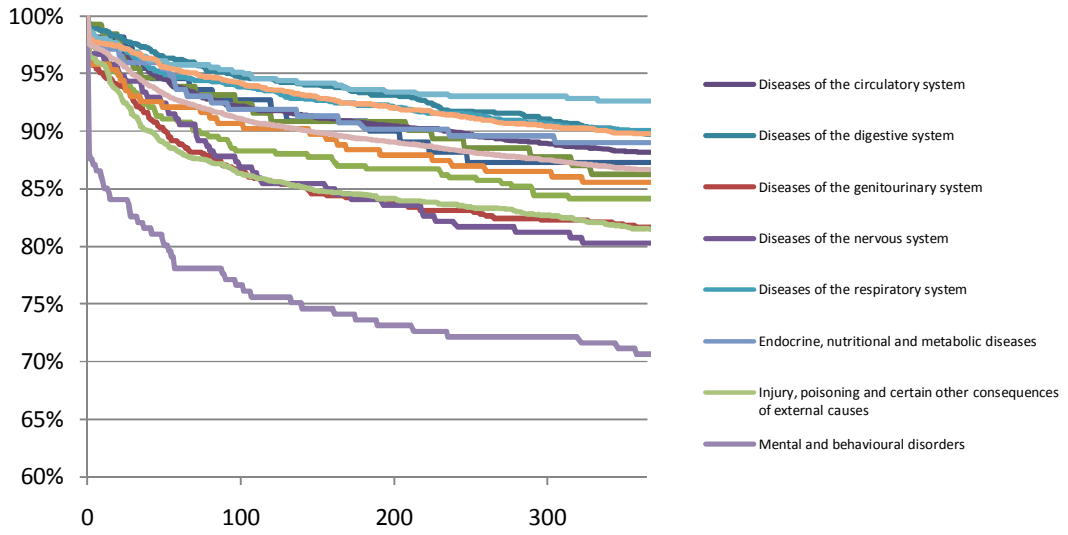


Figure 8.4 shows a similar example but comparing the patterns according to the principal diagnosis on admission to hospital. For this group of people aged 75 and above, the group admitted with a mental health problems show a sudden change shortly after discharge. Otherwise there is clear pattern between disease categories.

Figure 8.4 Analysis of a cohort of older people showing the proportion not receiving high-intensity social care after discharge from hospital following an emergency hospital admission (stratified according to diagnosis).



9. THE FUTURE AGENDA

In this section we explore some of the options to further exploit our work on linked health and social care data. We have grouped our comments under three headings: next steps for case finding; other applications of the linked data; and recognition of other research development issues. We hope that this will provide an agenda for debate.

PILOT IMPLEMENTATION APPROACHES

Once the information governance issues have been resolved, we would recommend that sites who wish to run the model should start by applying the models to historic data in order to identify a few patients from various strata of risk so that their needs can be identified. For example, a site might identify ten patients with a risk score of 90– 100; ten patients with a risk score of 75– 85; and ten patients with a risk score 50–60. We suggest that a case note review be conducted on each of these patients by examining their GP notes, pattern of hospital use and social care records, both for the two years prior to prediction and for the year following prediction. Ideally the patients and their carers would also be interviewed, although this is likely to be too time-consuming and might well require ethics approval. The information gathered could then be anonymised and reviewed by a multi-disciplinary team of doctors, social workers, nurses and allied health professionals. The aim of this team would be to recommend the types of 'upstream' care that might have made a difference in each case. These recommendations could then be used to help determine business cases for various types of service to be offered and help guide the choice of which strata of risk to target.

Finally, if a predictive model for social care were to be rolled out in practice then we would strongly recommend evaluating the cost-effectiveness of any interventions offered in terms of their impact on healthcare and social care use and costs. Such a development and research project might be similar to the Department of Health's Whole System Demonstrator pilots that are testing the cost-effectiveness of telehealth and telecare by means of a large-scale randomised controlled trial.

DEVELOPING POOLED COMPARABLE DATASETS

The health sector is able to benefit from the existence of national datasets with a standard structure and coding. Though far from perfect, these datasets find increasing use in the funding and research of health services. We note the absence of similar datasets for social care. Here any comparative data tends to be on an aggregated basis. The problem with aggregated data is that data collection only happens periodically (typically once a year) because the definitions have to be fixed and are slow to change. Furthermore, the degree of detail in terms of types of user, service or geography is inevitably limited. We therefore welcome the recent initiatives by the Information Centre to move toward a national approach for social care.

Within a period of about six months, our project team has been able to obtain, link and organise the health and social care data from five sites covering a population of over three million older people. Incidentally, we have also obtained social care data from three other areas (as yet only partially linked to healthcare data). We have been informally approached by some other local authorities who are keen for their data to be analysed in this way. We believe we have shown what is possible and it would be good to maintain momentum. One way to do this would be to establish a 'data laboratory' that would seek to collate linked, pseudonymous health

and social care data from consenting sites across the country. This dataset would then be available as a planning and research tool, both for the sites themselves and for others.

We also believe that there are opportunities to undertake some further data linkage and matching on specific topics. For example, it would in theory be possible to use information about the location of care homes as a way of indicating which people use care homes (including those who are paying for themselves). Thus, for example, hospital records that capture postcodes could be linked to care home addresses. However, such an application would require clarity on the rules for information governance.

PERSONALISATION AND BUDGET SETTING

The Department of Health is committed to developing models of care that give people more choice and control over the care that they receive, and it is promoting direct payments and individual budgets as a way of achieving this. Individuals receive either cash or a notional sum for them to spend on their care or support package in the way that best suits their own requirements.

Several government departments, including the Department of Health, Communities and Local Government and the Department for Work and Pensions, have been working to bring together different funding streams for each individual. Examples include 'Supporting People', Disabled Facilities Grant and 'Access to Work'. We believe that some of the approaches developed in this project could help with the setting of budgets – particularly preventive budgets – that span several funding organisations.

TRANSITIONS THROUGH CARE

The novel linked datasets that we have collated give us a brief glimpse of a range of analyses that might be possible. The work on trajectories (briefly described in Section 8) is an example of how these data might best be used to understand the typical pathways that people experience through care services. We are aware that our data could reveal important insights concerning:

- the probability of transitions between states of dependence
- the returns on investment based on comparisons with a realistic expectations drawn from other areas
- the extent to which service substitution between areas does or does not lead to long-term cost savings
- the potential scope for insurance-based funding schemes and some likely long-term costs scenarios.

DEVELOPING INTEGRATED INFORMATION SYSTEMS

This project has highlighted some of the insights that can be gained through the exploitation of existing datasets in ways that do not compromise user confidentiality. The ability to conduct this type of work should improve in the coming years as more health and social records go paperless and more social care records begin to record NHS numbers. The importance of high-quality information systems has been cited as a critical factor in the development of successful integrated care organisations.[32]

It is not difficult now to imagine a world where individuals' pseudonymous records are analysed in real-time, with recommendations, risk scores and quality gaps being 'pushed' to professionals within the electronic record, similar to the ways in which products are recommended by Amazon and other online retailers according to previous browsing and purchase histories.

A very simple yet powerful addition to electronic medical records and social care records would be a graphic representation of all the encounters of an individual with health and social care. Having this information available would help professionals place the current consultation in context.

The logical extension of this work would be to move towards person-based resource allocation across both health and social care. The Department of Health recently funded a piece of analysis that adopts a new approach to resource allocation by attributing notional needs for health to individuals. These needs' estimates are then aggregated for appropriate populations to estimate a required share of expenditure. The project focused on healthcare, indeed mainly acute healthcare. However, the same approach might potentially be used for resource allocation across health and social care.

ASSESSING IMPACTABILITY

Several predictive model vendors in the United States are developing 'impactability models' designed to identify the subgroup of at-risk patients who are most amenable to 'upstream' care. Some of the strategies being pursued for predicting impactability have important implications for tackling healthcare inequalities.[33]

Impactability models can be considered as a second step that refines the output of a predictive risk model (see Table 9.1).

Table 9.1 Types of impactability model

Type of impactability model	Details
Actionable conditions	Builds on the concept of ambulatory care sensitive conditions to rank all diseases in terms of how amenable they are to preventive care
Quality gaps	Ranks patients according to the number and nature of quality gaps. In a patient with diabetes, not taking cholesterol-lowering medication (for example, a statin) constitutes one gap
Likelihood to engage*	Ranks patients according to likelihood of engaging with 'upstream care'
Receptivity	Suggests how best to approach each individual patient in order to offer 'upstream care'

*In the United States, this type of model may be used to exclude patients who are deemed least likely to engage with 'upstream care', for example, people whose first language is not English, people with mental illness, alcoholism, substance abuse and single parents. In contrast, we would propose using this type of model to channel additional resources to these hard-to-reach groups thereby tackling healthcare inequalities. Source: GH Lewis[33]

NEXT STEPS

We wish to make the following specific recommendations that could help further this important agenda:

1. The predictive accuracy of our models is comparable with those of the models used by the NHS to predict hospital admissions. However, the ways that our models might be used in practice is less clear-cut. We suggest that it will be important to support their piloting and testing as case-finding tools for social care in the participating sites. Pilots like this could help answer key question such as:
    - choice of information/data to be used in modelling
    - thresholds for risk scores and relevant interventions
    - costs and benefits of implementation
    - range of interventions that might be triggered by a risk score and the evidence base for those interventions.

2. The linkage of health and social care data continues to raise questions about information governance processes. In order to exploit the huge potential of linked data there needs to be a better understanding of what is and is not permissible. We suggest that a clear protocol needs to be agreed by the National Information Governance Board for Health and Social Care and the Information Centre, and that this should be widely disseminated.

3. We believe there will be value in developing an experimental dataset that includes pseudonymous social care datasets linked with health data from a number of sites. Linked data will be useful in the evaluation of several initiatives, including:
   - impact of reablement according to risk profile of user
   - personalised budgets
   - integrated and coordinated care
   - eligibility criteria for a national care service.

4. Social care data systems would benefit from greater consistency in how they record and code information. Some work is underway by the Care Services Efficiency Delivery programme at the Department of Health and the NHS Information Centre with the development of the Tools for Rapid Integration of Public Submissions.[34] This has the potential to access a variety of data sources, including the main operational database systems, and convert them into a consistent social care database. This is a prerequisite for detailed, meaningful comparisons across areas. In turn, we expect that this will lead to improvements in coding and data quality for social care, especially if a secondary uses service (SUS) for social care were to be established analogous to the existing healthcare SUS and Hospital Episode Statistics.

5. Data linkage should be promoted for the commissioning of integrated health and social care organisations. The benefits of linked data should be assessed in a number of settings including:
   - information for professional and clinical staff
   - tools for planning and commissioning care.

REFERENCES

1. Cousins MS, Shickle LM and Bander JA (2002) 'An introduction to predictive modeling for disease management risk stratification', *Disease Management*, vol 5 (3): 157–167.

2. Stuck AE, Egger M, Hammer A, Minder CE and Beck JC (2002) 'Home visits to prevent nursing home admission and functional decline in elderly people: systematic review and meta-regression analysis', *JAMA*, vol 287, 1022–1028.

3. Lewis GH (2007) *Predicting Who Will Need Costly Care: How best to target preventive health, housing and social programmes*. London: The King's Fund.

4. Billings J, Dixon J, Mijanovich T and Wennberg D (2006) 'Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients', *BMJ*, vol 333: 327.

5. US Congress (1996) *Health Insurance Portability and Accountability Act*. Available at http://en.wikipedia.org/wiki/Health_Insurance_Portability_and_Accountability_Act

6. Integrated Research Application System. Available at https://www.myresearchproject.org.uk

7. NRES Ethics Consultation e-group (2006) 'Differentiating audit, service evaluation and research'. Available at www.nres.npsa.nhs.uk/EasySiteWeb/GatewayLink.aspx?alId=320

8. Patient Information Advisory Group (PIAG). Available at www.dh.gov.uk/ab/Archive/PIAG/index.htm

9. PIAG (2008) *The Use of Patient Information in the Long Terms Conditions Programme*. Available at www.advisorybodies.doh.gov.uk/piag/piag-ltc-oct2008.pdf

10. DH Care Services Efficiency Delivery. Available at http://webarchive.nationalarchives.gov.uk/+/www.dh.gov.uk/en/SocialCare/Socialcarereform/Careserviceseffi ciencydelivery/index.htm

11. DH Care Services Efficiency Delivery. Available at http://webarchive.nationalarchives.gov.uk/+/www.dh.gov.uk/en/SocialCare/Socialcarereform/Careserviceseffi ciencydelivery/index.htm

12. NHS Strategic Tracing Service. Connecting for Health. Available at www.connectingforhealth.nhs.uk/systemsandservices/nsts. Accessed 2 November 2008.

13. Pope G, Kautter J, Ellis R, Ash A, Ayanian J, Iezzoni L, Ingber M, Levy J and Robst J (2004). 'Risk adjustment of Medicare capitation payments using the CMS-HCC Model', *Health Care Financing Review*, 25 (4), 119–141.

14. NHS. Accident and Emergency Hospital Episode Statistics. Available at www.ic.nhs.uk/statistics-and-data-collections/hospital-care/accident-and-emergency-hospital-episode-statistics-hes

15. Wennberg D, Siegel M, Darin B, Filipova N, Russell R, Kenney L, Steinort K, Park T, Cakmakci G, Dixon J, Curry N and Billings J (2006). *Combined Predictive Model: Final report and technical documentation*. Available at www.kingsfund.org.uk/document.rm?id=6745

16. Weiner JP, Starfield BH, Steinwachs DM and Mumford LM (1991) 'Development and application of a population-oriented measure of ambulatory care case mix', *Medical Care*, 29 (5):453–472.

17. NHS Information Service (2008) *Frequently Asked Questions on the Home Help/Home Care Collection (HH1)*. Available at www.ic.nhs.uk/webfiles/Services/Social%20care/Collections/2008%20to%2009/HH1/FAQs%20HH1%202008-09%20v2.pdf

18. NHS Information Service *General Guidance Notes for Completion of the SR1 Return (2005/06)*. Available at www.ic.nhs.uk/webfiles/Services/Social%20care/Collections/2005%20to%2006/sr1guid.pdf

19. Experian Mosaic™ UK 2009. Available at www.experian.co.uk/www/pages/what_we_offer/products/mosaic_uk.html

20. Altman D and Bland J (1994) 'Statistics notes: diagnostic tests 1: sensitivity and specificity', *BMJ*, 308:1552

21. Szmukler G (2001) 'The mathematics of risk assessment for serious violence', *Psychiatric Bulletin*, 25: 359. Available at http://pb.rcpsych.org/cgi/content/full/25/9/359

22. Wennberg D, Siegel M, Darin B, Filipova N, Russell R, Kenney L, Steinort K, Park T, Cakmakci G, Dixon J, Curry N and Billings J (2006) *Combined Predictive Model: Final report and technical documentation.* Available at www.kingsfund.org.uk/document.rm?id=6745

23. Wildman MJ, Sanderson C and others (2009) 'Predicting mortality for patients with exacerbations of COPD and Asthma' *Thorax*, 64: 128–132.

24. Harrell FE, Lee KL and Mark DB (1996) 'Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors', *Statistics in Medicin*e, 15: 361–387.

25. Stuck AE, Egger M, Hammer A, Minder CE and Beck JC (2002) 'Home visits to prevent nursing home admission and functional decline in elderly people: systematic review and meta-regression analysis', *JAMA*, 287: 1022–1028.

26. Brodaty H, Gresham M and Luscombe G (1997) 'The Prince Henry Hospital dementia caregivers' training programme', Dementia study, *International Journal Geriatric Psychiatry*, 12: 183–192.

27. PIAG (2008) *The Use of Patient Information in the Long Terms Conditions Programme*. Available at www.advisorybodies.doh.gov.uk/piag/piag-ltc-oct2008.pdf

28. NHS Wales. Delivering PRISM. Available at www.wales.nhs.uk/sites3/page.cfm?orgid=770&pid=41418

29. Milliman (2009) *Milliman Care Guidelines*. Available at www.milliman.com/expertise/healthcare/products-tools/milliman-care-guidelines/. Accessed 8 February 2009.

30. Weber C and Neeser K (2006) 'Using individualised predictive disease modelling to identify patients with the potential to benefit from a disease management program for diabetes mellitus,' *Disease Management*, 9, no. 4:242–255.

31. www.walsall.gov.uk/index/reablement-2.htm

32. Rosen R and Ham C (2008) *Integrated Care: Lessons from evidence and experience*. London: Nuffield Trust.

33. Lewis GH (2010) ' "Impactibility models": identifying the subgroup of high-risk patients most amenable to hospital-avoidance programs,' *Milbank Quarterly,* 88, no. 2: 240-255.

34. CIPFA (2010) *Social Care Panel Newsletter.* Available at www.cipfa.org.uk/newsletters/socialcare/download/SCP_summer2010.pdf (p8).

The Nuffield Trust
FOR RESEARCH AND POLICY
STUDIES IN HEALTH SERVICES

Predicting social care costs: a feasibility study

Predictive models have the potential to provide a better experience for service users and to offer more cost-effective care. Such models are increasingly being used in health care to identify people at high risk of unplanned hospital admission, so that preventive care can be effectively targeted. This report presents the findings of research into whether such models could also be used in social care to predict an individual person's future need for intensive social care.

In addition to the predictive models developed, the work also generated important lessons about the potential of linked health and social care to support policy analysis and to guide the planning and commissioning of services.

Predicting social care costs: a feasibility study will be of interest to health and social care policy-makers, senior managers and practitioners, and others involved in commissioning, as well as academics and students in the fields of health care and social policy.

This report forms part of the Trust's work on commissioning of health care. Further information on our work in this area can be found at: **www.nuffieldtrust.org.uk/projects/**

Nuffield Trust
59 New Cavendish Street
London W1G 7LP
Email: info@nuffieldtrust.org.uk
www.nuffieldtrust.org.uk