COMBINED PREDICTIVE MODEL

FINAL REPORT & TECHNICAL DOCUMENTATION

DECEMBER 2006

NHS
DH Department of Health
NYU – Robert F. Wagner Graduate School of Public Service
King's Fund
Health Dialog UK Limited

ACKNOWLEDGEMENTS

The following individuals are the principal contributors to the development of the Combined Predictive Model:

HEALTH DIALOG
David Wennberg, MD, MPH
Matt Siegel
Bob Darin
Nadya Filipova, MS
Ronald Russell, MS
Linda Kenney
Klaus Steinort
Tae-Ryong Park, PhD
Gokhan Cakmakci

KING'S FUND
Jennifer Dixon, MBChB, PhD
Natasha Curry

NEW YORK UNIVERSITY
John Billings

We would like to acknowledge the invaluable support and participation of numerous organisations involved in this project including its funders, the Department of Health and Essex Strategic Health Authority (acting on behalf of all 28 Strategic Health Authorities), as well as the National Health Service staff who joined the project steering group. We would also like to thank the Croydon and South Warwickshire Primary Care Trusts for supplying the data used in the development of the Combined Predictive Model, as well as the Tower Hamlets and Southwark Primary Care Trusts for their data collection efforts.

CONTENTS

IDENTIFYING RISK ALONG THE CONTINUUM

To meet national goals for reductions in emergency bed days and effective administration of practice-based commissioning, National Health Service (NHS) organisations have highlighted the need for tools to assess patient needs across the continuum of care. A risk stratification tool called the Combined Predictive Model (the Combined Model) has been developed to provide a rich segmentation of patients at each section of the continuum. The model is based on a comprehensive dataset of patient information, including inpatient (IP), outpatient (OP), and accident & emergency (A&E) data from secondary care sources as well as general practice (GP) electronic medical records.

Stratification results derived from the Combined Model are shown in Figure 1. The model was developed on a 50% random sample of data from two Primary Care Trusts (PCTs) and validated on the other 50% random sample.* All patients in the validation sample were ranked based on their risk for emergency admission and placed into segments. Relative utilisation rates are shown for patients in each segment for the ***year following prediction*** compared to average utilisation rates across the entire population. For example, patients in the top 0.5% predicted risk segment were 18.6 times more likely than the average patient to have an emergency admission in the year following prediction.

Through identifying relative risk along the continuum, the Combined Model allows NHS organisations to develop and tailor intervention intensity to match the expected 'returns'.

Through identifying relative risk along the continuum, the Combined Model allows NHS organisations to develop and tailor intervention intensity to match the expected 'returns'. Previously, this level of detail and stratification were unavailable to the NHS, but the Combined Model allows for development and implementation of these strategies across patient segments.

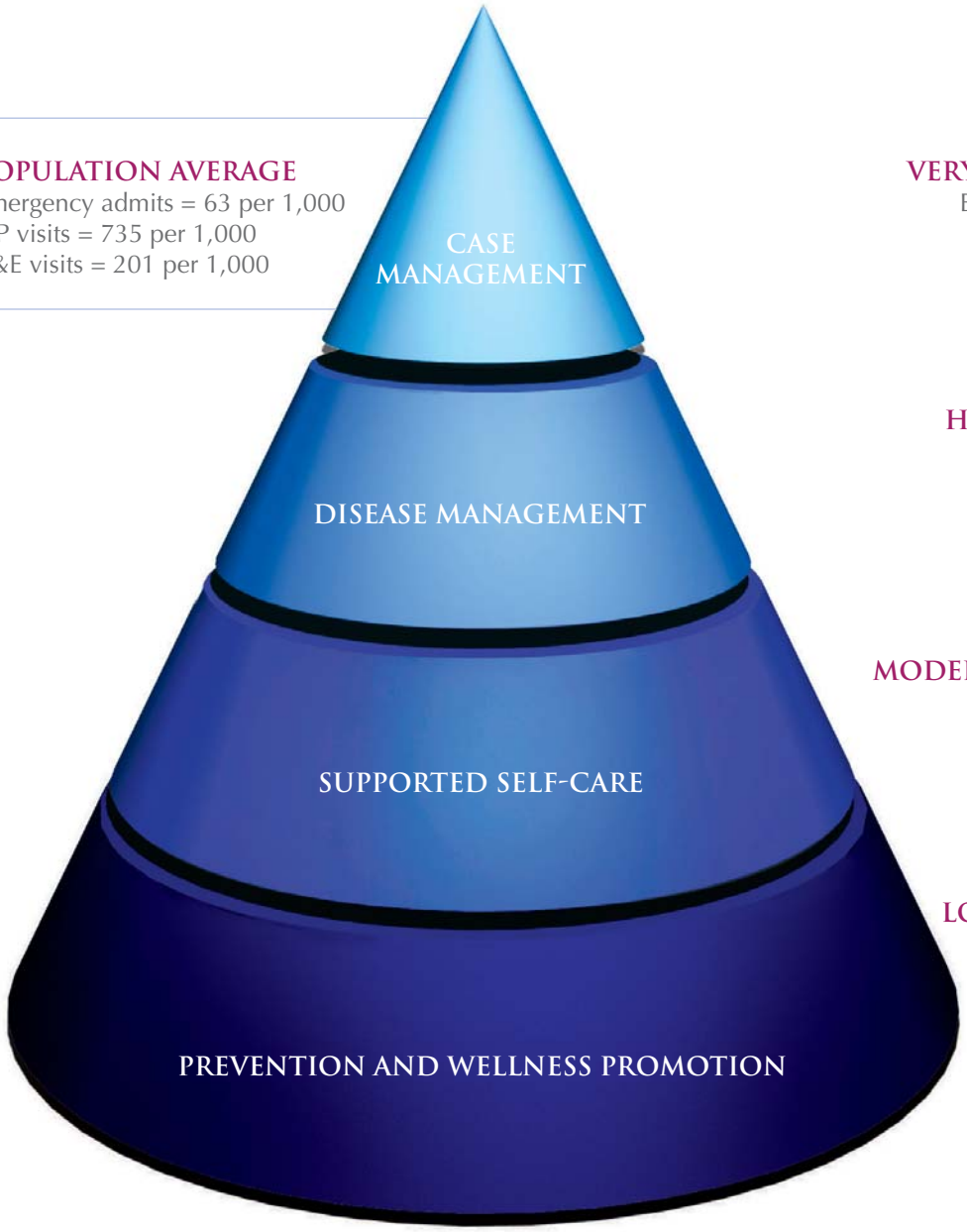
The ability to tailor interventions to expected risk based on stratification results such as these is critical for three reasons. First, Practice-based Commissioning will require that clinicians and managers use resources wisely, particularly given available supply of care management interventions. Second, while much of the current intervention focus is on the tip of the pyramid, need is distributed along the continuum. Third and most important, we recognize that more care is not always necessarily wanted or needed. A generic intervention model applied to all patients within a practice would likely increase utilisation among those at the bottom of the pyramid.[1-3]

*Analyses in the Final Report are based on validation of the Combined Model on a random 50% sample of the total population of the two PCTs which provided data for its development. The validation analyses were based on the time period of 1 April 2002-31 March 2004 to predict emergency admissions in the following 12 months of 1 April 2004-31 March 2005. More information on methodology is included in Appendix A.

FIGURE 1

SEGMENTATION OF PATIENT POPULATION USING COMBINED MODEL

POPULATION AVERAGE
Emergency admits = 63 per 1,000
OP visits = 735 per 1,000
A&E visits = 201 per 1,000



VERY HIGH RELATIVE RISK 0.5%
Emergency admits = 18.6 x average
OP visits = 5.8 x average
A&E visits = 8.5 x average

HIGH RELATIVE RISK 0.5 - 5%
Emergency admits = 5.5 x average
OP visits = 3.8 x average
A&E visits = 2.9 x average

MODERATE RELATIVE RISK 6 - 20%
Emergency admits = 1.7 x average
OP visits = 1.9 x average
A&E visits = 1.4 x average

LOW RELATIVE RISK 21 - 100%
Emergency admits = 0.5 x average
OP visits = 0.6 x average
A&E visits = 0.8 x average

Source - Health Dialog UK

BACKGROUND

THE NEED FOR PREDICTIVE CASE FINDING

The development of long term conditions management, including case management, is becoming established across England. These efforts have been 'encouraged' by the release of various national strategic papers; a national Public Service Agreement target has been set to improve outcomes for people with long term conditions. This agreement calls for a personalised care plan for vulnerable people most at risk, and includes as a goal the reduction of emergency bed days by 5% by March 2008.

Case finding is essential for effective long term conditions management. Predicting who is most at risk of emergency admissions is a critical function of case finding.

Case finding is essential for effective long term conditions management. Predicting who is most at risk of emergency admissions is a critical function of case finding. Tools that can identify those who can most benefit from outreach and targeted interventions require a high degree of accuracy to ensure that there is a match between intervention intensity and risk.

To address this need, a package of predictive case finding algorithms has been commissioned by the Department of Health (DH)/Essex Strategic Health Authority from a consortium of the King's Fund, New York University and Health Dialog. This consortium has developed three tools. The first two are aimed at identifying Patients At Risk for Re-hospitalisation *(PARR1 and PARR2)*. PARR1 uses data on prior hospitalisations for certain 'reference conditions' to predict risk of re-hospitalisation while PARR2 uses data on any prior hospitalisation to predict risk of re-hospitalisation. The third tool is aimed at identifying risk along the continuum (the *Combined Model*). The PARR models use IP data only, while the Combined Model supplements these data with OP, A&E and GP data. The Combined Model was developed with two PCTs which supplied the data for its development.

THE PATIENTS AT RISK FOR RE-HOSPITALISATION (PARR) MODEL AND CASE MANAGEMENT

PARR1 and PARR2, tools that identify very high risk patients, have been previously released. Both use inpatient data to produce a 'risk score' showing a patient's likelihood of re-hospitalisation within the next 12 months. Risk scores range from 0 – 100, with 100 being the highest risk.

Since their release in Autumn 2005, the PARR algorithms have been widely distributed and shown to be effective in identifying patients with high utilisation of secondary care services[4]. These patients are being targeted for intervention by Community Matrons, Virtual Wards and other similar case management approaches. Given the limited data set used to identify these patients and the resulting narrow population targeted when looking only at re-admissions, the need for additional tools exists to identify patients across a broader spectrum of care needs and levels of intervention.

The need for additional tools exists to identify patients across a broader spectrum of care needs and levels of intervention.

THE COMBINED MODEL

THE COMBINED MODEL AND SEGMENTATION STRATEGIES

To meet this broader need and to determine whether the addition of further data sets improves predictive accuracy, a third algorithm has been developed which combines secondary care data with GP electronic records. This Combined Model is able to:

- **Improve predictive accuracy for very high risk patients**
- **Predict risk of hospital admission for those patients who have not experienced a recent emergency admission**
- **Stratify risk across all patients in a given health economy to help NHS organisations understand drivers of utilisation at all levels**

The ability to identify emerging risk patients will enable NHS organisations to take a more strategic approach to their care management interventions. For example, PCTs will be able to design and implement interventions and care pathways along the continuum of risk, ranging from:

- **Prevention and wellness promotion for relatively low risk patients**
- **Supported self-care interventions for moderate risk patients**
- **Early intervention care management for patients with emerging risk**
- **Intensive case management for very high risk patients**

The broad application of the Combined Model will allow segmentation of an entire population into relative risk segments and facilitate matching the intensity of outreach and intervention with the risk of unwarranted secondary care utilisation.

The broad application of the Combined Model will allow segmentation of an entire population into relative risk segments and facilitate matching the intensity of outreach and intervention with the risk of unwarranted secondary care utilisation. The ability to apply the intervention in a targeted fashion increases the likelihood that patients will receive the care they want (and nothing more) and the care they need (and nothing less).

WHAT DOES THE COMBINED MODEL DO?

The aim of the Combined Model is to use a broader and more comprehensive set of data to identify patients who may become frequent users of secondary care services. Through prospectively identifying these patients, the appropriate levels of outreach and intervention can be applied; from helping patients at lower risk to manage their conditions with information and self-management support, to providing intensive case management support for patients at the highest levels of risk.

The Combined Model was developed using a split sample methodology on data from two PCTs with a total population of 560,000. Details of the development methodology and population can be found in Appendix A. The model takes primary and secondary care data for an entire patient population and stratifies those patients based upon their risk of emergency admission in the next 12 months. With access to this broader set of data beyond just inpatient data, the Combined Model is not limited to identification of very high risk patients based solely on past admissions. The Combined Model offers a tool to help design, commission and implement an overall long term conditions programme strategy.

The Combined Model offers a tool to help design, commission and implement an overall long term conditions programme strategy.

THE COMBINED MODEL

THE COMBINED MODEL ENHANCES PREDICTIVE ACCURACY

In addition to stratifying an entire patient population and identifying emerging risk, the Combined Model is also effective in identifying patients in the very high and high risk segments of the population. In the sections below, we discuss the clinical and utilisation profiles of patients who fall into these segments, highlighting the opportunities for impact. In the highest risk segments where the most intensive outreach will be targeted, such as case management interventions, the Combined Model improves predictive performance over the PARR (i.e., PARR2) model for the same populations. Figure 2 below shows the Positive Predictive Value (PPV)* for different cuts of population size identified by either the Combined or the PARR model.

*PPV is a reflection of the number of patients who actually had an emergency admission in the year following prediction out of all of the patients who were predicted to have an emergency admission within that segment. For example, 586 out of the top 1000 patients predicted by the Combined Model actually had an emergency admission in the year following prediction as compared with 505 out of the top 1000 PARR patients.

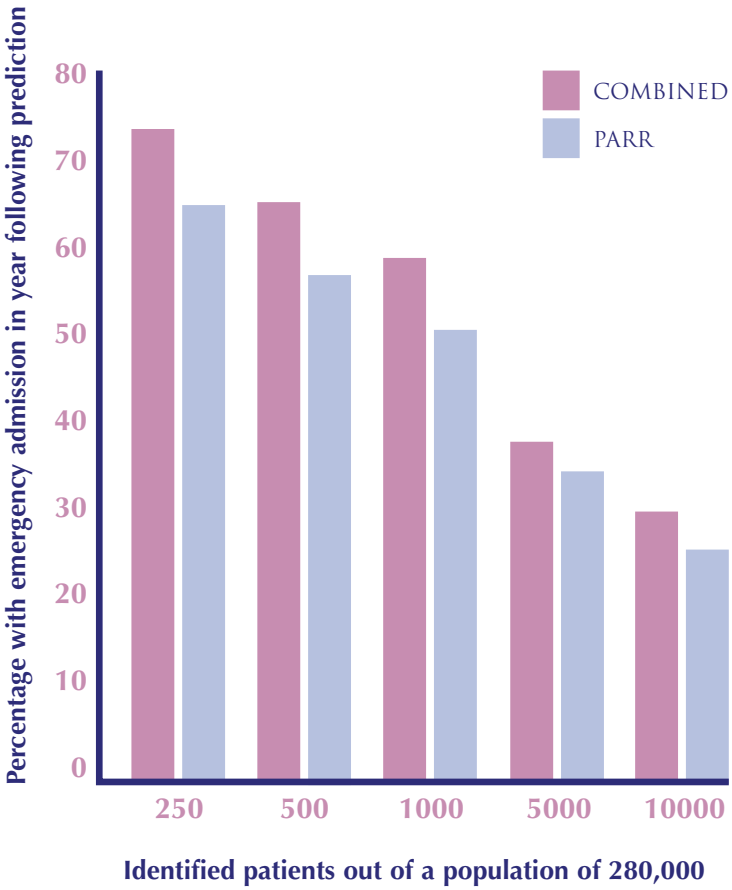


FIGURE 2

POSITIVE PREDICTIVE VALUE FOR COMBINED MODEL VS. PARR

GENERAL PRACTICE DATA ADD TO THE PREDICTIVE ACCURACY

The Combined Model was also developed to determine whether GP practice data add to predictive accuracy compared to the PARR model and against models that might include outpatient attendances and A&E data but not GP data. Figure 3 below shows the PPVs for different risk segments, still within the very high and high risk categories, for the Combined Model compared to the Combined Model with GP variables removed (i.e., using IP, A&E, and OP data only) and also compared to the PARR model. Comparing the full Combined Model against the Combined Model without the GP data included in the prediction allows one comparison of the relative impact of including GP data.

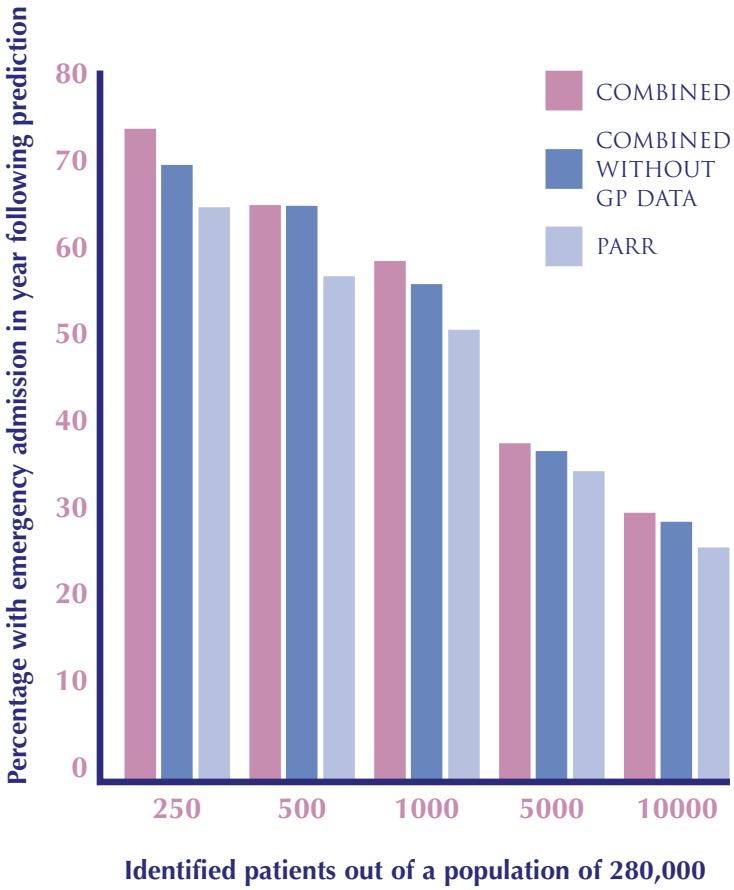


FIGURE 3

PPV FOR COMBINED MODEL WITH AND WITHOUT GP DATA VS. PARR

THE COMBINED MODEL

As highlighted on page 11, Figure 3 shows that a more inclusive model using inpatient, outpatient, and A&E data alone outperforms PARR, and the full Combined Model which also includes GP data outperforms both models at almost all risk segments.

With the additional predictive accuracy achieved by introducing the OP, A&E, and GP data sets, the 'break even' analysis of the potential cost savings that can be achieved is enhanced when compared with PARR, particularly when identifying very high risk patients. Figure 4 below shows scenarios built by running the Combined Model and PARR2 on the validation sample and focusing only on the segments where case management interventions might be most suitable. An intervention cost of £500 per patient and intervention impact of 20% is assumed. The additional predictive accuracy of the Combined Model allows PCTs to design interventions with greater potential for net cost savings.

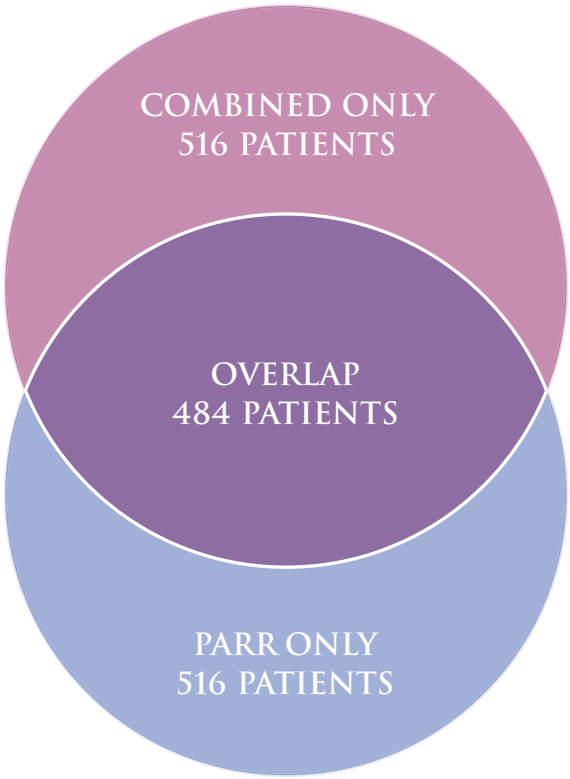
FIGURE 4

BREAK-EVEN FOR VERY HIGH RISK PATIENTS

Risk Score Cut-off		Number of true positives	Number of false positives	Cost per patient	Total intervention cost	Emergency admissions within 12 months per true positive	Estimated impact of intervention	Estimated cost per admission	Total intervention savings	Net savings or loss
Top 250	Combined Model	184	66	£500	£125,000	2.85	20%	£2,100	£220,248	£95,248
	PARR	163	87	£500	£125,000	2.75	20%	£2,100	£188,265	£63,265
Top 500	Combined Model	327	173	£500	£250,000	2.54	20%	£2,100	£348,844	£98,844
	PARR	285	215	£500	£250,000	2.71	20%	£2,100	£324,387	£74,387
Top 1000	Combined Model	586	414	£500	£500,000	2.33	20%	£2,100	£573,460	£73,460
	PARR	505	495	£500	£500,000	2.44	20%	£2,100	£517,524	£17,524

THE COMBINED MODEL INTRODUCES A NEW PATIENT POPULATION

The PARR and Combined Models identify different patients, even at the highest risk levels. The Venn diagram in Figure 5a below demonstrates the overlap between the PARR and Combined Models using the top 1,000 patients as an example: those patients who are identified in PARR only, those identified in the Combined Model only, and those identified in both models.



With the additional predictive accuracy achieved by introducing the OP, A&E, and GP data sets, the 'break even' analysis of the potential cost savings that can be achieved is enhanced when compared with PARR, particularly when identifying very high risk patients (as shown in Figure 4 on page 12).

FIGURE 5A

VENN DIAGRAM OF PARR AND COMBINED MODEL PATIENT POPULATIONS OUT OF TOP 1,000 IDENTIFIED

THE COMBINED MODEL

Figure 5b below shows the Combined Model patients at different cut points stratified into emerging risk patients, including both patients who have no prior inpatient admission history (light blue), as well as patients who have an admission history but a lower risk score from the PARR model (dark blue) and those identified by PARR (purple). For example, out of the 1000 highest risk patients identified in the Combined Model sample, approximately 48% of them would also have been identified in the top 1000 patients using PARR in the same sample. Forty seven percent of the top 1000 would have been identified in PARR but would have a relatively lower risk score. A further 5% would not have been identified at all using PARR.

The addition of patients who would have been missed by PARR altogether, due to lack of prior inpatient admissions, and patients who would have been identified at much lower risk levels by PARR, due to its reliance on inpatient data only, is significant. The Combined Model's use of richer data sets allows for risk stratification at levels conducive to more effective early intervention as it identifies patients before they have deteriorated to the point of multiple inpatient admissions.

The addition of patients who would have been missed by PARR altogether, due to lack of prior inpatient admissions, and patients who would have been identified at much lower risk levels by PARR, due to its reliance on inpatient data only, is significant.

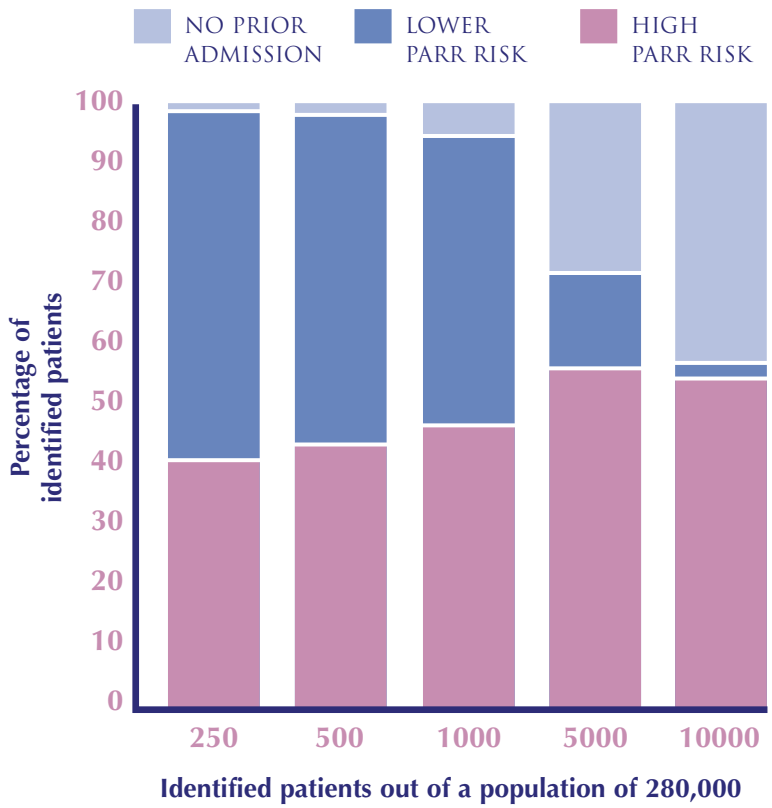


FIGURE 5B

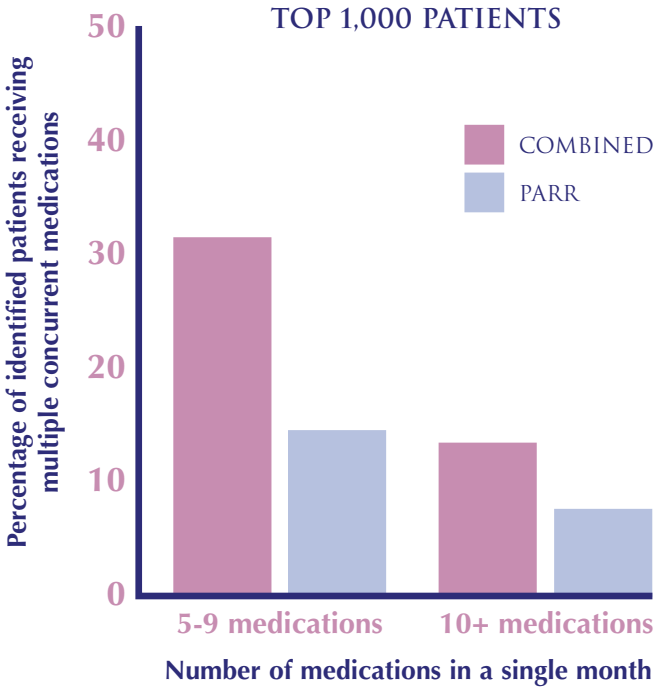
OVERLAP OF PATIENTS IDENTIFIED BY COMBINED MODEL AND PARR

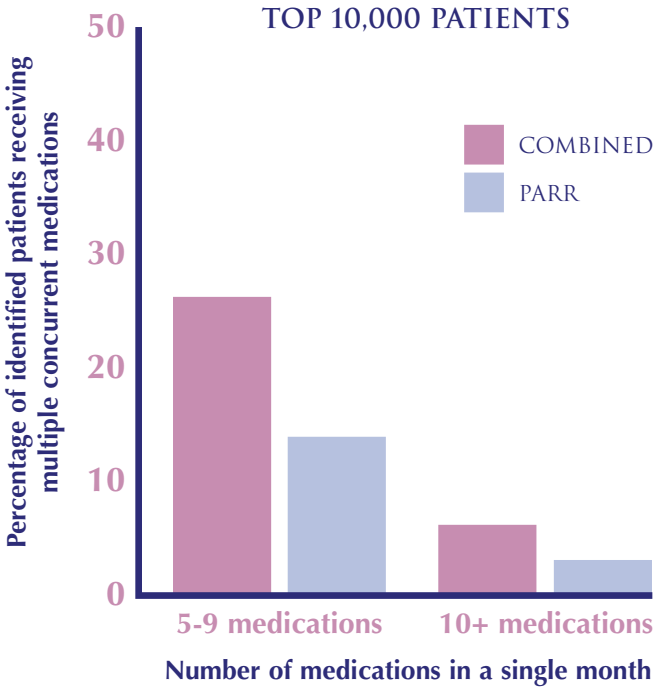
THE COMBINED MODEL IDENTIFIES PATIENTS WITH RICH CLINICAL PROFILES AND OPPORTUNITIES TO IMPACT FUTURE UTILISATION AND CLINICAL CARE

The addition of GP, OP, and A&E data sources in the Combined Model gives further clinical insights into the status of identified patients and the factors that are contributing significant risk for emergency admission. In addition, the clinical profile that emerges from creating the input data required to implement the Combined Model provides a much more descriptive clinical roadmap of how to tailor the intervention to the needs of the patients identified.

FIGURES 6A & 6B

POLYPHARMACY UTILISATION AMONG PATIENTS IDENTIFIED BY COMBINED MODEL AND PARR





Across a number of different measures, the high risk patient population being identified by the Combined Model is rich in opportunities for impact. For example, Figures 6a and 6b above show the percentage of patients in the top 1,000 and top 10,000 identified in the Combined Model and PARR that are taking between five and nine and 10 or more different prescription drugs in a single month.

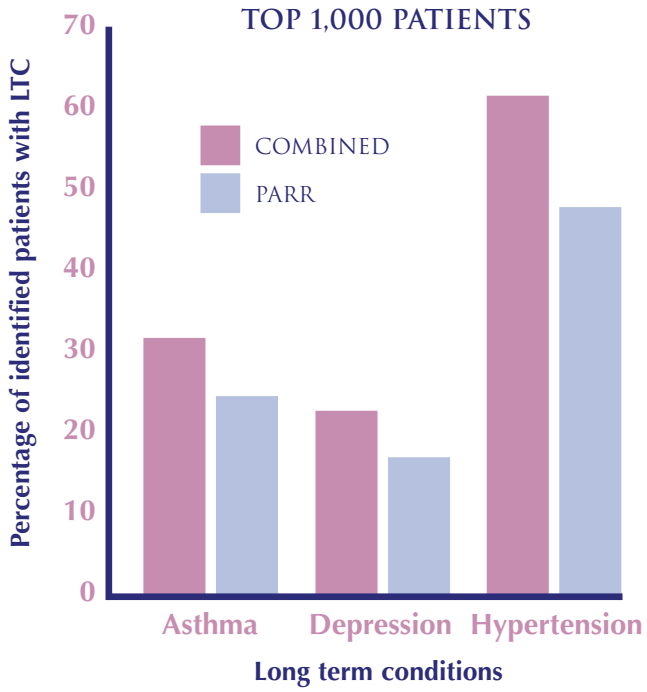
THE COMBINED MODEL

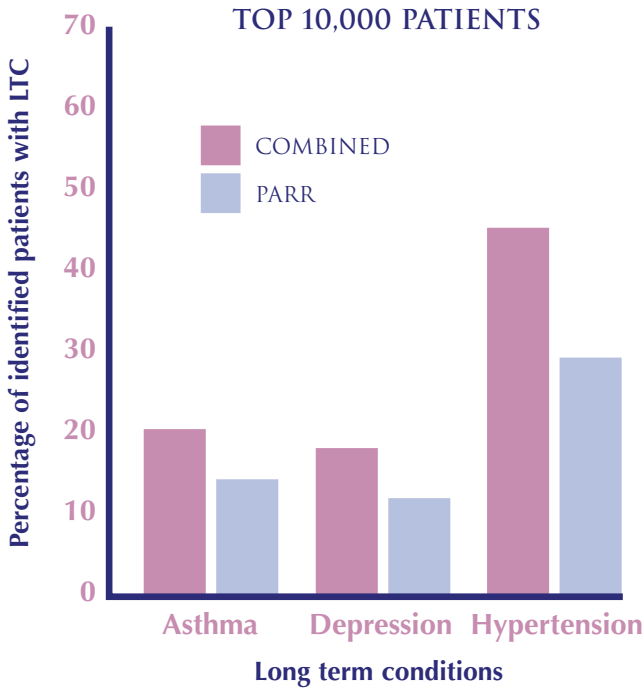
Polypharmacy issues are a significant area of focus for high intensity and/or telephonic interventions; the Combined Model identifies a set of patients with higher rates of polypharmacy-related concerns than the PARR model. This clinical information, only available through the linking of the different data sets, will have a direct impact on the type and intensity of intervention design planned, such as the use of pharmacy experts to look at polypharmacy issues and how to manage those for improved outcomes and lower cost.

Figures 7a and 7b below look at the prevalence of key chronic diseases in the top 1,000 and top 10,000 patients (a summary of clinical profile variables across cutpoints is shown in Appendix B). The prevalence of impactable conditions such as asthma, depression, and hypertension is higher in the top 1,000 Combined Model patients than the top 1,000 PARR patients.

FIGURES 7A & 7B

LONG TERM CONDITION PREVALENCE AMONG PATIENTS IDENTIFIED BY COMBINED MODEL AND PARR





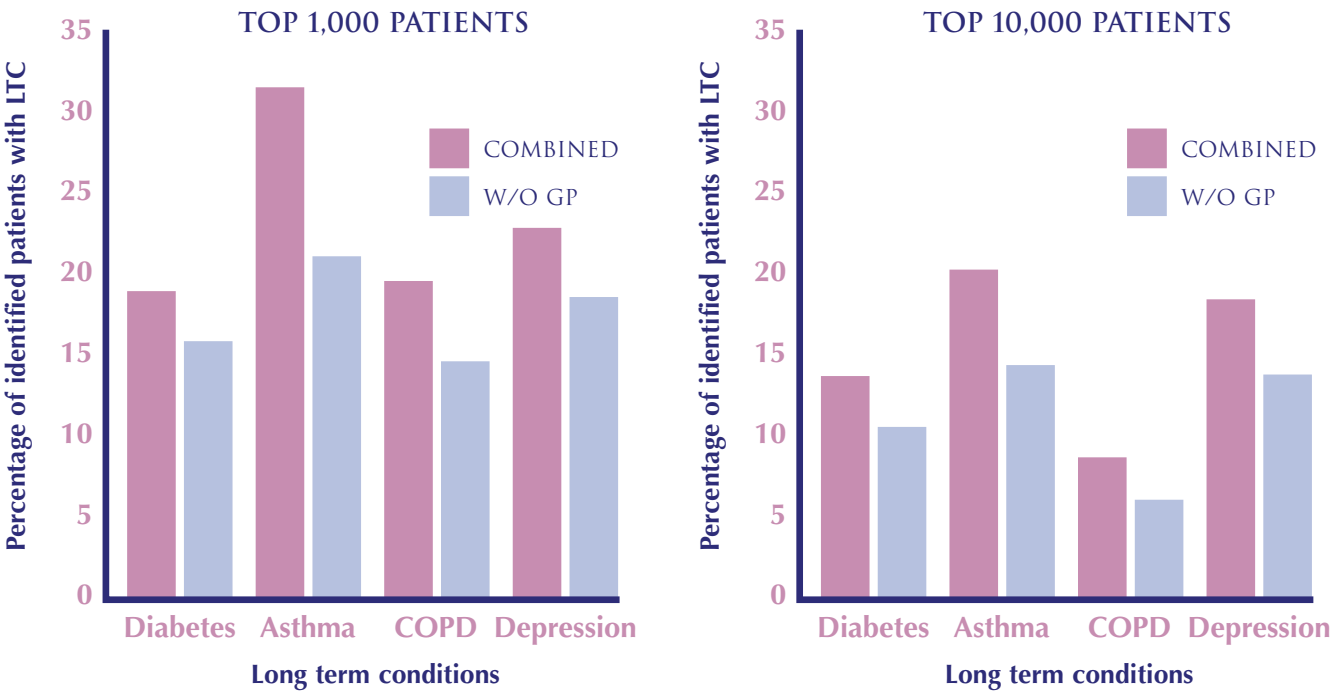
This is consistent with the different patient population being introduced using the more comprehensive data set and the ability to identify emerging risk patients. Similarly, for the top 10,000 patients, the Combined Model is consistently identifying patients with higher long term condition prevalence and more impactable opportunities across all conditions (as shown in Appendix B).

GENERAL PRACTICE DATA HELPS IDENTIFY PATIENTS WITH IMPACTABLE CONDITIONS

Including GP practice data, in addition to the secondary care data, significantly enhances the opportunity to identify patients with long term conditions and the overall richness of the clinical opportunities for intervention. Figures 8a and 8b below show the prevalence of diabetes, asthma, chronic obstructive pulmonary disease (COPD) and depression within both the top 1,000 and top 10,000 patient segments when comparing the Combined Model with and without GP variables. Adding GP data enhances the ability of the model to identify more patients with impactable long term conditions.

FIGURE 8A & 8B

LONG TERM CONDITIONS IN COMBINED MODEL AND COMBINED MODEL EXCLUDING GP VARIABLES



As the true goal of these segmentation efforts is to reduce the risk rather than describe it, the additional clinically relevant information in the GP practice data is essential for carrying out interventions across patient segments. Additional clinical profile information for the Combined Model and Combined Model excluding GP variables is included in Appendix B.

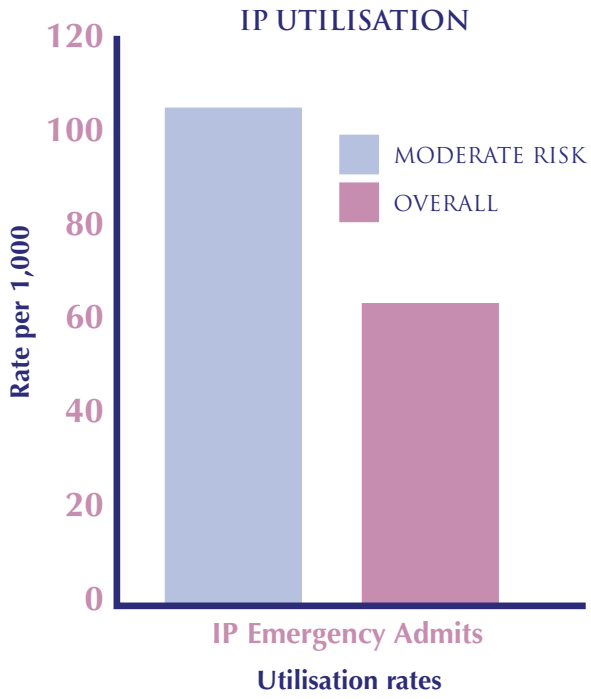
THE COMBINED MODEL

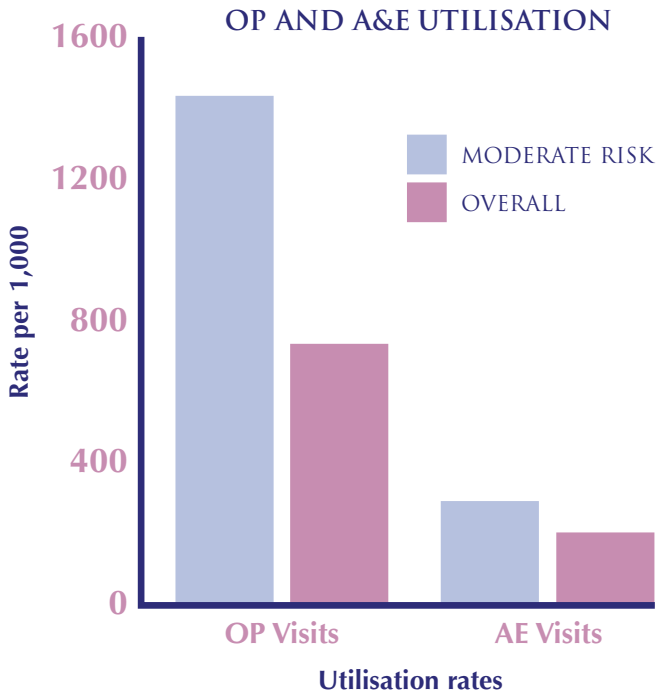
THE COMBINED MODEL OFFERS THE ABILITY TO IDENTIFY OPPORTUNITIES IN OTHER SEGMENTS OF THE RISK PYRAMID

As discussed earlier, the Combined Model identifies patients across the continuum of risk. This allows NHS organisations to tailor targeted outreach and intervention to the relative risk of individual patients in each segment of the risk pyramid (shown on page 5). Most of this document focuses on those in the very high and high risk segments where case management and disease management interventions involving direct interaction with patients may be warranted. However, there are also opportunities to design lower intensity strategies for supported self-care for patients in the moderate risk segment (6-20%) such as support via telephone, mail, internet, text messaging and/or email.

FIGURES 9A & 9B

UTILISATION OF MODERATE RISK PATIENTS IN RISK PYRAMID





As Figures 9a and 9b above demonstrate, there is ample secondary care utilisation driving cost within this segment of more than 40,000 patients that could be addressed using lower-intensity interventions. Patients in the moderate risk segment have nearly twice as many outpatient attendances, 70% more emergency admissions, and 40% more A&E attendances when compared with the average person in the population.

Figure 10 below demonstrates that there is also significant clinical opportunity within the moderate risk group. For example, compared with population averages, patients in the moderate risk segment are more than twice as likely to have polypharmacy utilisation of between five and nine different drugs in a single month. In addition, there is relatively high prevalence of impactable long term conditions in this segment which, if unmanaged, may lead to patients progressing up the pyramid. For example, hypertension prevalence in this group is 18% compared with 9% in the overall population.

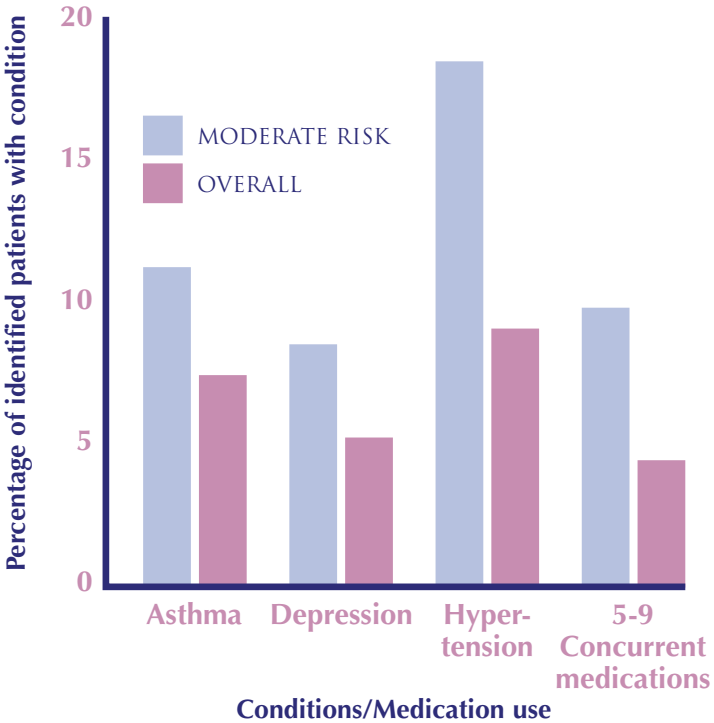


FIGURE 10

CLINICAL PROFILE OF MODERATE RISK PATIENTS IN RISK PYRAMID

CONCLUSION

The Combined Model significantly extends the range of opportunities for the NHS and clinicians interested in long term condition interventions. The ability to stratify the entire population allows PCTs to develop intervention strategies aimed at reducing immediate, intermediate and longer term risk. Further, the addition of rich, clinical detail allows PCTs to not only segment populations, but also begin to assemble the clinical interventions for each of the different segments.

The Combined Model offers an increase in predictive power for the highest risk patients, and also facilitates the identification of a much broader population with emerging risk.

The findings from the Combined Model show that it holds significant potential value for NHS organisations seeking to develop population-based strategies for utilisation reduction and quality improvement. Whilst the PARR model has offered the NHS a nationwide tool that allows for quick identification of the very highest risk patients, it has been limited to identifying only those individuals at the highest end of the risk pyramid. The Combined Model offers an increase in predictive power for the highest risk patients, and also facilitates the identification of a much broader population with emerging risk. An integrated approach, using both tools, which matches interventions of varying intensity to population needs across the continuum of risk levels will be an essential component of PCTs' care management strategies, and the Combined Model offers an important set of tools for PCTs to design and implement these strategies.

REFERENCES

1 Fisher, Wennberg, Stukel, et al, "The Implications of Regional Variations in Medicare Spending. Part 1: The Content, Quality, and Accessibility of Care," Annals of Internal Medicine, 2003; 138:273-287.

2 Fisher, Wennberg, Stukel, et al, "The Implications of Regional Variations in Medicare Spending. Part 2: Health Outcomes and Satisfaction with Care," Annals of Internal Medicine, 2003; 138:288-298.

3 O'Connor, Llewellyn-Thomas, and Flood, "Modifying Unwarranted Variations In Health Care: Shared Decision Making Using Patient Decision Aids," Health Affairs – Web Exclusive, 2004; VAR 63-72.

4 Billings, Dixon, Mijanovich, and Wennberg, "Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients," British Medical Journal, 2006; 333:327-330.

APPENDIX A

SUMMARY OF DATA SOURCES AND METHODOLOGY

The Combined Model was developed on a total population of 560,000 patients from two PCTs using three years of hospital data (April 2002 – March 2005), including inpatient (IP), outpatient (OP), and accident and emergency (A&E) attendance data. Additionally, primary care data for the same time period were included from the two PCTs, including lab, diagnosis, and encounter information from general practices within those PCTs. Unfortunately, pharmacy data were only included for one of the two PCTs that supplied the primary care data. In addition, social services data were requested from the PCTs participating in the Combined Model development work. Health Dialog was able to link the social services information to the clinical data supplied for only one of the PCTs and only in a very small percentage of patients due to complications with the data. This proportion of linked social and clinical service records at the patient-level was not sufficient for inclusion in the Combined Model.

The model was developed using logistic regression on a random selection of 50% of the available data (known as the 'development sample'). Data for the period of April 2002 through March 2004 were mined for predictor variables associated with risk of admission during the time period of April 2004 through March 2005. The model was validated by applying the variable beta weights resulting from the logistic regression analyses to the remaining 50% of data (known as the 'validation sample'). All Combined Model results shown in the Final Report are for this validation sample only and are compared with PARR scores for patients from the same validation sample and same time period.

In development, more than 850 variables were considered for inclusion. These variables included a combination of values from administrative records and derivations from those values. Derived variables included proxy variables for long-term conditions (drawn from GP and IP encounters), polypharmacy (drawn from Read codes evaluated on a monthly basis), and changes in lab values (derived from GP encounters). Each variable was also coded into five mutually exclusive time periods to account for recency of occurrence and patterns of recurrence. Each variable was assessed independently for its relationship with inpatient emergency admission before being included in a multivariate model.

APPENDIX B

CLINICAL PROFILE A

CLINICAL PROFILE INFORMATION FOR PATIENTS IDENTIFIED AT DIFFERENT RISK SEGMENTS USING THE COMBINED MODEL VERSUS PATIENTS IDENTIFIED AT THE SAME RISK SEGMENTS USING THE PARR MODEL

		LONG TERM CONDITION											
Identified patients	Model	Asthma	COPD	Depression	Diabetes	Hypertension	Cancer	CHD	CHF	Avg. age	5 - 9 Medications	10+ Medications	Avg. length of stay*
10,000	Combined	20.1	8.2	17.9	13.7	45.2	14.9	15.5	6.5	67.3	26.0	6.0	11.4
	PARR	14.3	5.1	11.8	9.3	29.0	9.4	12.0	4.6	55.4	14.1	3.0	10.4
5,000	Combined	23.3	11.0	20.6	16.2	51.4	15.3	18.5	9.4	69.7	29.5	8.6	11.5
	PARR	16.6	8.8	13.7	12.9	38.7	15.1	18.8	8.3	66.2	17.6	4.3	11.0
1,000	Combined	31.4	19.6	22.7	18.8	61.2	16.4	24.5	15.9	71.6	31.5	13.4	10.3
	PARR	24.3	20.6	16.9	19.9	47.7	20.0	24.8	19.2	69.5	14.6	7.4	9.9
500	Combined	34.0	22.6	25.2	19.2	63.2	16.4	28.8	19.4	70.7	30.6	14.6	9.8
	PARR	28.4	24.6	16.4	23.0	49.0	21.0	28.8	24.4	69.1	14.8	8.6	9.4
250	Combined	40.8	28.4	24.4	21.2	62.8	17.6	29.6	23.2	69.5	32.0	18.0	8.9
	PARR	30.8	26.8	20.0	22.8	52.4	20.8	30.8	29.2	68.8	14.4	10.0	9.8

* per emergency admission

CLINICAL PROFILE B

CLINICAL PROFILE INFORMATION FOR PATIENTS IDENTIFIED AT DIFFERENT RISK SEGMENTS USING THE COMBINED MODEL VERSUS PATIENTS IDENTIFIED AT THE SAME RISK SEGMENTS USING THE COMBINED MODEL BUT WITH GP VARIABLES EXCLUDED (FROM PREDICTION)

		LONG TERM CONDITION											
Identified patients	Model	Asthma	COPD	Depression	Diabetes	Hypertension	Cancer	CHD	CHF	Avg. age	5 - 9 Medications	10+ Medications	Avg. length of stay*
10,000	Combined	20.1	8.2	17.9	13.7	45.2	14.9	15.5	6.5	67.3	26.0	6.0	11.4
	w/o GP	14.1	6.2	13.8	10.5	34.6	14.6	13.5	6.0	67.2	17.3	4.0	11.3
5,000	Combined	23.3	11.0	20.6	16.2	51.4	15.3	18.5	9.4	69.7	29.5	8.6	11.5
	w/o GP	15.8	8.0	15.4	12.5	40.0	15.1	15.9	8.4	69.6	19.6	5.3	11.3
1,000	Combined	31.4	19.6	22.7	18.8	61.2	16.4	24.5	15.9	71.6	31.5	13.4	10.3
	w/o GP	21.0	14.5	18.2	15.7	49.1	17.6	22.4	16.0	71.7	20.9	8.2	10.5
500	Combined	34.0	22.6	25.2	19.2	63.2	16.4	28.8	19.4	70.7	30.6	14.6	9.8
	w/o GP	25.8	18.0	20.6	17.4	55.6	17.0	26.6	19.6	71.7	22.6	10.6	10.3
250	Combined	40.8	28.4	24.4	21.2	62.8	17.6	29.6	23.2	69.5	32.0	18.0	8.9
	w/o GP	27.2	21.6	20.0	18.4	56.8	19.2	28.0	24.4	70.9	22.4	11.6	8.8

* per emergency admission

COMBINED PREDICTIVE MODEL

TECHNICAL DOCUMENTATION

The purpose of this document is to describe the specific data collection, management, and analysis procedures used to develop, adjust, and apply the Combined Predictive Model (the Combined Model) as defined in the document below. This document is supported by the electronic files contained in the eMedia Appendices. They provide details on code resolution, data encryption and logical groupings of critical categories. eMedia Appendices are referenced in appropriate places throughout this document. Although the Combined Model was developed by Health Dialog using the SAS programming language, the same procedures can be implemented in any procedural programming language (e.g. Basic, C, SPSS, STATA). Implementation of the Combined Model requires a familiarity with fundamental programming skills involving database creation, analysis and processing.

Statistical modelling is inherently dependent upon the data used to create that particular model. The intention of this document is to provide detail to highlight distributions of data used in the development of the Combined Model, to note exceptions as necessary, and to define the steps required for NHS organisations to implement the Combined Model.

DATA EXTRACTION, ASSESSMENT, AND TRANSFORMATION SUMMARY

Optimal implementation of the Combined Model requires a minimum of two years of historical data to predict admissions for the following year. Production of the model and review of the results requires a three-year database (the first two years to implement the model and predict risk and the third year to evaluate anticipated vs. actual results). The Combined Model was developed from three years of hospital records, incorporating inpatient (IP), outpatient (OP), and accident and emergency (A&E) data. Additional data were collected from Primary Care Trust (PCT)-affiliated general practices, including drug and Read code information. Authorisation to obtain general practice (GP) data is critical to successfully creating a robust data set from which to perform predictive analysis.

Five PCTs supplied data for use in the predictive modelling process; three PCTs supplied GP data, one of which submitted only one practice worth of GP data for the requested years of April 2002 to March 2005. To preserve critical patient confidentiality as part of security agreements, all personally identifiable information was submitted in encrypted form.

Once PCT authorisation had been obtained, extracted data were encrypted by either the PCT or the King's Fund, and then delivered to the King's Fund office in London via CD, portable hard drive, or secure courier. Authorisation was also obtained to use hospital data which was a simpler process due to the fewer authorisations required than was the case for the sharing of GP and social services (SS) data. While all PCTs supplied available IP, OP, and A&E data, only three of the recruited PCTs supplied GP data (with one only able to supply one practice worth of data), and only two supplied SS data. Only data for which authorisation was obtained was utilised for analysis.

Critical concerns regarding the data centered on issues of consistency of encrypted data and availability. In order to consolidate data for patients existing within different subsets, the National Health Service (NHS) identification number had to be non-reversibly encrypted in consistent fashion to protect the true identity of the patient but still provide a link between patient data sources in order to develop patient-level clinical profiles. Certain critical elements required for analysis were derived from a multitude of sources where available, such as age at time of encounter and gender, available in some datasets and not in others due to encryption. Identification of the minimal elements not incorporating personally identifiable data elements yet supporting analysis would greatly ease the implementation of future work.

Initial assessment of the data revealed some challenges, particularly with regard to missing data. For example, age was missing from two sources of data, but was processed from other sources to determine age at the time of encounter. In all cases, where specified, date of birth was encrypted, challenging the identification of age at time of encounter when not specified.

Invalid codes were removed from diagnosis and procedure codes. Codes were also normalised between different versions of coding for consistency of analysis. After transformation, data were consolidated into a central data warehouse for processing and analysis.

EXTRACTION

Primary Care Trusts supplied data relating to IP admissions, OP encounters, and A&E visits. For the purposes of analysis, one year of data was considered to extend from April 1st to March 31st of the following year. PCTs were requested to supply data with as much information as available in each of these areas, not necessarily limited to Hospital Episode Statistics (HES) or other standardised data sets. As part of established security agreements, personally identifiable information was supplied in an encrypted format. Sex and age (or year of birth) were requested to be supplied in all data sets, although some sources encrypted these in addition to other personally identifiable information.

The King's Fund worked with appropriate parties at the recruited PCTs to acquire data. Once authorisation was obtained from each PCT, the data were extracted and encrypted where necessary by the PCT or the King's Fund, and delivered to the King's Fund in London. Health Dialog and the King's Fund used an encryptor tool that used an NHS-approved Secure Hash Algorithm (SHA-1) to encrypt fields for the NHS number, post code, and date of birth. These fields were used to cross-reference the data sets and maintain consistent associations between them. Data encryption required a significant amount of time due to the large size of the data sets and the design of the original tool.[1]

Once encrypted, the data sets were transferred to Health Dialog on CD or portable hard drive by authorised Health Dialog employees or secure courier. Stored data were secured to limit access physically and electronically to authorised individuals only.

IP, OP, and A&E data were obtained from records available to the PCT. GP data were extracted from GP information systems, centrally collected, and provided to the PCT; or retrieved through individual queries and extracts.

Initial results from the data extraction are summarised below:

PCT	IP	OP	AE	GP	SS
1	3 years	3 years	3 years	3 years	3 years[2]
2	3 years	3 years	3 years	3 years	3 years[3]
3	3 years	3 years	3 years	2 years[4]	Not Available
4	3 years	3 years	3 years	Not Available	Not Available
5	3 years	3 years	1 year	Not Available	Not Available

1 Inconsistent encryption keys and shifting underlying data syntax introduced problems in some data sets, requiring reprocessing of queries and encryption tasks.
2 Not linkable, due to absence of NHS numbers
3 Not linkable, due to absence of NHS numbers
4 Single practice only

ASSESSMENT

Data were received in text files, in comma- or tab-separated value format, and read into SAS data sets. One PCT provided hospital data in Access databases. An initial check was made to verify the following:

- Files were readable, without corrupt records
- Data matched the identified layout
- Data were provided for the time period requested
- Data sets were provided with consistently encrypted NHS numbers for association
- SS data could be linked on key encrypted elements
- Age and sex were available and consistent on all data sets.

In a few cases, the data received were corrupted or incomplete. Indicators of corruption included stray characters in fields known to be numeric, misaligned values, and excessive numbers of incorrect values for fields associated with known national codes. These data were resent after re-extraction and encryption.

NHS numbers were not present in some of the data; in the absence of other key identifiers, these data were removed from consideration. When NHS numbers were missing, key fields that were encrypted in the same fashion were used for matching, if possible. However, this resulted in few matching records, and non-matching records were removed from consideration. For example, up to 14% of A&E data were missing an NHS number. Because a blank NHS number still yielded an encrypted value, care was taken to correctly identify the blank NHS number through its encrypted value and remove those records from the data set.

The unavailability of age or year of birth information in several data sets posed a major challenge. Only one PCT for which the age was missing was able to re-run the extraction of data. For other PCTs, some age information was available in certain subsets of the data, but not in others. In these cases, Health Dialog matched the records based on ages from the known data sets, developing a cross-reference list of known age by patient where available.

One of the two sets of SS data could not be associated with other data sets from the same PCT due to apparent problems in the syntax of encrypted fields. An alternative to the NHS number (which is absent from SS data) — matching on encrypted date of birth, encrypted post code, and gender from SS data to other data sets — yielded a very low match rate. Because encryption occurred on the alphabetical representation of the date of birth, differences in that format between data sets rendered matching impossible. Obtaining a consistent format for dates between data sets was critical. For example, the encrypted values for “30-APR-2000” would yield a different value than the encrypted value for “30/04/2000”. Post code was also difficult to match, since some systems used a dashed format, and others used a compressed post code.

Despite known issues concerning data quality, for practical purposes it was assumed that no better data were available - data issues were documented and a work-around devised wherever possible. For example, age was matched for individuals common to more than one data set. A reverse look-up was performed for the birth date assigned for that individual (to the nearest year), based on matching age to encrypted birth date, and assigning the same age adjustment for similar records with the same encrypted value. Simply put, matching patients of known ages with a derived birthdate to the nearest year can be used to identify the year of birth by matching encrypted birthdates.

The table below summarises the records received from each PCT, and the distribution of records by year:

	Year	IP		OP		A&E		GP	
PCT1	2001	1,064	0.5%	0	0.0%	0	0.0%	0	0.0%
	2002	74,305	31.9%	432,626	25.4%	90,166	28.5%	4,964,198	29.2%
	2003	77,241	33.1%	790,920[5]	46.4%	105,416	33.3%	6,109,899	36.0%
	2004	80,528	34.5%	480,895	28.2%	121,080	38.2%	5,907,834	34.8%
	2005	0	0.0%	0	0.0%	0	0.0%	0	0.0%
	Other	70	0.0%	0	0.0%	0	0.0%	7,624	0.0%
	Total	**233,208**		**1,704,441**		**316,662**		**16,989,555**	
PCT2	2001	636	0.2%	0	0.0%	0	0.0%	-	-
	2002	64,777	24.3%	193,408	24.2%	627	29.5%	-	-
	2003	69,609	26.1%	198,438	24.9%	706	33.2%	-	-
	2004	71,461	26.8%	203,323	25.5%	792	37.3%	-	-
	2005	0	0.0%	203,004	25.4%	0	0.0%	-	-
	Other	60,542	22.7%	7	0.0%	0	0.0%	-	-
	Total	**267,025**		**798,180**		**2,125**		**-**	
PCT3	2001	1,064	0.5%	0	0.0%	-	-	-	-
	2002	61,268	30.8%	179,749	30.9%	-	-	-	-
	2003	66,805	33.6%	195,564	33.7%	-	0.0%	-	0.0%
	2004	69,573	35.0%	205,561	35.4%	-	0.0%	-	0.0%
	2005	0	0.0%	0	0.0%	-	0.0%	-	0.0%
	Other	56	0.0%	0	0.0%	-	0.0%	-	0.0%
	Total	**198,766**		**580,874**		**-**		**-**	
PCT4	2001	1,994	0.3%	117	0.0%	14,750	11.0%	2,406,669	10.5%
	2002	195,581	32.1%	395,167	29.8%	34,298	25.6%	3,124,660	13.6%
	2003	191,523	31.4%	452,156	34.1%	36,108	27.0%	2,704,438	11.8%
	2004	202,900	33.3%	465,301	35.1%	48,768	36.4%	2,694,195	11.8%
	2005	17,160	2.8%	13,522	1.0%	0	0.0%	483,660	2.1%
	Other	41	0.0%	35	0.0%	0	0.0%	11,483,770	50.2%
	Total	**609,199**		**1,326,298**		**133,924**		**22,897,392**	
PCT5	2001	1,056	0.4%	0	0.0%	0	0.0%	-	0.0%
	2002	85,810	32.4%	222,425	28.9%	43,658	24.1%	-	0.0%
	2003	79,773	30.1%	258,247	33.6%	60,346	33.3%	-	0.0%
	2004	91,706	34.6%	275,710	35.9%	76,568	42.3%	-	0.0%
	2005	6,857	2.6%	12,138	1.6%	393	0.2%	-	0.0%
	Other	9	0.0%	1	0.0%	0	0.0%	-	0.0%
	Total	**265,211**		**768,521**		**180,965**		**-**	

Some PCT submissions included extra data, outside of the window of evaluation – these values tended to be small and were excluded from further consideration.

Wide variations exist between different PCTs with regard to both prevalence and robustness of data. In order to help isolate and prevent potential problems, it is prudent to execute checks to evaluate the data. These checks include prevalence, adherence to standard coding schemes, and expected demographics regarding population composition. Included in this document are the recommended data layouts for each data set, which were also used to develop the predictive models.

The data sets used in the modelling were based on the SAS data read during the initial data receipt and quality check. Initially, as all data elements available were to be considered in the model, layouts were not standardised but included all data sent.

5 Includes duplicate data – duplicates later removed, resulting in 462,632 records.

INPATIENT (IP) DATA

LAYOUT – IP DATA

In order to ensure a consistent framework for modelling, data were generated according to a common and consistent format, as outlined below. A 'Y' under the column 'Required for Model' indicates variables which are needed to generate the model parameters.

Field Name	Definition	Required for Model?
Admin_Category	Type of patient (NHS, private, etc)	
Admission_Date	Date of admission	**Y**
Age_at_Start_of_Episode	Age in years	**Y**
Chronically_Sick_or_Disabled	Indicates if the patient is disabled or chronically sick	
Cons_Specialty_Code	Specialty of consultant	
Consultant_Code	GMC code that identifies the consultant	
Date_of_Birth	Date of birth	
Date_of_Death	Date of death	
Date_of_Decision_to_Admit	This date may be the same as the date of admission (e.g. most emergency admissions). Alternatively, a decision can be made to admit at a future date. This decision denotes that the PATIENT is intended to be admitted to a hospital bed, either immediately or subsequently in the future. It records the event that a clinical decision to admit a PATIENT to a hospital bed has been made by or on behalf of someone, who has the right of admission to a hospital provider for that patient.	
Date_of_Primary_Procedure	This is the Clinical Intervention Date of the PRIMARY OPERATION (OPCS-4)	
Destination_on_Discharge_Code	This records the destination of a PATIENT on completion of the Hospital Provider Spell. It can also indicate that the PATIENT died or was a still birth	
Detention_Category	Mental category (mental illness, mental impairment, etc)	
Diag_Version	This is used in the Clinical Information Group of the CDS to denote the scheme basis of a Diagnosis	
Discharge_Date	The date a PATIENT was discharged from a Hospital Provider Spell	**Y**
Discharge_Method_Code	The method of discharge from a Hospital Provider Spell	**Y**
District_of_Residence_Code	District where the patient resides	
Electoral_Ward	Electoral ward of patient residence	
Episode_End_Date	The date that an Episode ends	
Episode_ID	Unique identifier for this episode	
Episode_Number	Episode order - first episode in a spell = 1	
Episode_Start_Date	Date episode starts	
Ethnic_Origin	The new national code must be entered as the first character in the 2-character field. The second character is an optional field only required for use locally. It must, however, be able to be grouped consistently with the 16 main categories.	
Hospital_Provider_Spell_No	A number to provide a unique identifier for each Hospital Provider Spell for a Health Care Provider	
HRG_Code	The National Schedule of Reference Costs, developed by the Department of Health, uses Healthcare Resource Groups (HRGs) as the basis for costing in-patient and day case services	
HRG3_Code	The 3 digit HRG assigned to the event	**Y**
Intended_Management_code	This categorisation describes what is intended to happen to the PATIENT	
Last_in_Spell_Indicator	This derived data element identifies whether the consultant episode is the final episode in the Hospital Provider Spell	

Field Name	Definition	Required for Model?
Legal_Status	The classification is required for all patients who have a Hospital Provider Spell which includes the care of a CONSULTANT in the psychiatric specialties or have been discharged from such a Hospital Provider Spell and are required to receive supervised aftercare under the provisions of the Mental Health (Patients in the Community) Act 1995	
Method_of_Admission_Code	The method of admission to a Hospital Provider Spell	Y
NHS_Number	NHS number if known	Y
Orig_Decided_to_Admit_Date	The date of the first decision to admit a patient to a Health Care Provider	
Patient_Classification_Code	A coded classification of PATIENTS who have been admitted to a Hospital Provider Spell	Y
Postcode_of_Usual_Address	The code allocated by the Post Office to identify a group of postal delivery points	
Practice_Code	The GP practice where the patient is registered	
Primary_Diagnosis	The main condition treated or investigated during the relevant episode of health care where there is no definitive diagnosis, i.e., the main symptom, abnormal findings, or problem (ICD-10)	Y
Secondary_Diagnosis[1..5]	Secondary or additional diagnosis	Y
Primary_Operative_Procedure	OPCS4	
Procedure_Status	OPERATION STATUS (procedure status) should be used once for each record to record states of knowledge regarding the operative procedure	
Procedure_Version	Procedure scheme in use. This is used in the Clinical Activity Group of the CDS to denote the scheme basis of an Intervention, Operation or A&E Treatment.	
Provider_ID	Facility providing the services	Y
Purchaser_ID	Usually a PCT identifier	
Record_Type	Indicates if the record for the episode is completed	
Referrer_Code	The ID of the health care professional making the referral	
Sex	Sex of patient	Y
Source_of_Admission_Code	The coded source of admission to a Hospital Provider Spell or a Nursing Episode when the PATIENT is in a Hospital Site or a Care Home	
Specialty_Code	This is the specialty in which the CONSULTANT is contracted or recognised	
Ward_Type	Ward of census	

ASSESSMENT – IP DATA

Each data element likely to be involved in the modelling process was examined, and distributions evaluated. (For a detailed listing of codes and their associated meanings, please refer to the six .csv files contained in the eMedia\Dictionary\IP folder.) Constraining data from each PCT to include only those records for the interval of 1 April 2002 through 31 March 2005, the following counts were observed:

PCT	Received	Within Interval	Duplicates	Remaining
PCT1	233,208	232,074	1,582	230,492
PCT2	267,025	205,847	7,700	198,147
PCT3	198,766	197,646	2,658	194,988
PCT4	609,199	590,004	2,267	587,737
PCT5	265,211	257,289	3,262	254,027
Total	**1,566,383**	**1,482,860**	**17,469**	**1,465,391**

Duplicated records which had identical information relating to the required fields identified above were removed.

Age_at_Start_of_Episode

In the data submission from the five PCTs, the following age distributions were noted for each PCT, with an admission date from 1 April 2002 through 31 March 2005:

Age	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
0-14	26,820	11.6%	12,155	6.1%	18,971	9.7%	31,557	5.4%	32,398	12.8%
15-44	83,076	36.0%	45,454	22.9%	72,826	37.3%	174,092	29.6%	97,966	38.6%
45-64	50,314	21.8%	47,909	24.2%	40,726	20.9%	132,442	22.5%	50,471	19.9%
65-74	29,189	12.7%	37,722	19.0%	24,312	12.5%	80,916	13.8%	35,467	14.0%
75+	41,084	17.8%	54,904	27.7%	28,928	14.8%	131,414	22.4%	37,725	14.9%
Invalid	9	0.0%	0	0.0%	0	0.0%	31,933	5.4%	0	0.0%
Missing	0	0.0%	3	0.0%	9,225	4.7%	5,383	0.9%	0	0.0%
Total	**230,492**		**198,147**		**194,988**		**587,737**		**254,027**	

This distribution reveals that substantial differences exist between PCTs with regard to the age reported and the groupings of that age. For example, PCT2 and PCT4 have a significantly larger number of patient records falling within the 75+ age group, suggesting either a difference in reporting, or an underlying difference in the population. Differences in distribution can affect the predictive nature of the derived model, and additional analysis can be required to identify the sources of those differences.

HRG3_Code

Although different versions of HRG coding were in use, Version 3.1 was provided by each PCT most often while the use of other versions appeared to be sporadic. This field was used for modelling because it provided the greatest number of fully specified HRG codes. Different sources also used one or more variant versions of the HRG coding scheme. Subsequent attempts to assign a uniform higher level HRG using the HRG grouper software available from the NHS were inconsistent with HRG codes on record. PCT3 did not submit HRG codes, so evaluation could not be performed on this variable for that PCT.

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
Unknown	17	0.0%	4	0.0%	0	0.0%	183	0.0%	1	0.0%
Missing	165,802	71.9%	3,613	1.8%	194,988	100.0%	3,211	0.5%	14,436	5.7%
Valid	64,673	28.1%	194,530	98.2%	0	0.0%	584,343	99.4%	239,590	94.3%
Total	**230,492**		**198,147**		**194,988**		**587,737**		**254,027**	

Method_of_Admission_Code

The method of admission is the national code assigned to a hospital provider spell.

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
Accident and emergency or dental casualty department	78,364	34.0%	30,170	15.2%	58,588	30.0%	132,202	22.5%	91,684	36.1%
Admitted ante-partum	22,953	10.0%	11,916	6.0%	18,979	9.7%	49,852	8.5%	27,883	11.0%
Admitted post-partum	429	0.2%	256	0.1%	277	0.1%	1,036	0.2%	243	0.1%
Baby born outside of this health care provider	21	0.0%	26	0.0%	410	0.2%	115	0.0%	24	0.0%
Bed Bureau	187	0.1%	15	0.0%	365	0.2%	472	0.1%	121	0.0%
Booked	51,919	22.5%	19,217	9.7%	40,762	20.9%	100,892	17.2%	31,238	12.3%
Consultant clinic, of this or another health care provider	3,237	1.4%	2,662	1.3%	3,076	1.6%	5,720	1.0%	4,894	1.9%
GP after a request for immediate admission has been made	929	0.4%	33,331	16.8%	1,604	0.8%	78,365	13.3%	5,131	2.0%

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
Not applicable	12	0.0%	15	0.0%	302	0.2%	5	0.0%	0	0.0%
Not known	125	0.1%	214	0.1%	1,747	0.9%	34	0.0%	5	0.0%
Other means, including admitted from the A&E department	5,462	2.4%	3,828	1.9%	4,973	2.6%	18,539	3.2%	3,319	1.3%
Planned	28,653	12.4%	44,371	22.4%	31,809	16.3%	85,757	14.6%	38,344	15.1%
The birth of a baby in this health care provider	14,632	6.3%	6,609	3.3%	12,633	6.5%	24,591	4.2%	19,473	7.7%
Transfer of any admitted patient from other hospital	3,515	1.5%	6,094	3.1%	1,099	0.6%	9,095	1.5%	5,903	2.3%
Waiting list	20,016	8.7%	39,494	19.9%	18,076	9.3%	80,627	13.7%	25,758	10.1%
Missing	17	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Unknown	21	0.0%	0	0.0%	288	0.1%	435	0.1%	7	0.0%
Total	**230,492**		**198,218**		**194,988**		**587,737**		**254,027**	

Discharge_Method_Code

The method of discharge code is the national code assigned at the time of discharge from the hospital provider spell. Values of “4” and “5” were used as exclusionary values, as they indicate that the patient was either deceased or a still birth.

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
Not applicable; hospital provider spell not yet finished	1,885	0.8%	4,792	2.4%	5,653	2.9%	1,541	0.3%	320	0.1%
Not known; a validation error	87	0.0%	3	0.0%	686	0.4%	26	0.0%	0	0.0%
Patient died	6,249	2.7%	5,759	2.9%	3,739	1.9%	19,597	3.3%	5,498	2.2%
Patient discharged by mental health review tribunal	19	0.0%	1	0.0%	28	0.0%	20	0.0%	2	0.0%
Patient discharged him/herself or was discharged by a relative	1,986	0.9%	857	0.4%	2,030	1.0%	3,893	0.7%	2,467	1.0%
Patient discharged on clinical advice	219,302	95.1%	186,784	94.2%	182,790	93.7%	560,210	95.3%	244,284	96.2%
Stillbirth	141	0.1%	20	0.0%	60	0.0%	71	0.0%	142	0.1%
Missing	768	0.3%	0	0.0%	0	0.0%	2,363	0.4%	0	0.0%
UNKNOWN	55	0.0%	2	0.0%	2	0.0%	16	0.0%	1,314	0.5%
Total	**230,492**		**198,218**		**194,988**		**587,737**		**254,027**	

Patient_Classification_Code

Patient classification is derived from the method of admission, intended management, and the duration of stay of the patient. For the purposes of modelling, only those records with a classification of 1 (ordinary admission) were included as an indicator of regular admissions.

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
Day case admission	55,173	23.9%	54,113	27.3%	43,258	22.2%	172,358	29.3%	58,840	23.2%
Missing	7	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Mothers and babies using only delivery facilities	98	0.0%	11	0.0%	29	0.0%	562	0.1%	1,000	0.4%
Not applicable	4	0.0%	0	0.0%	3	0.0%	1	0.0%	3	0.0%
Ordinary admission	161,483	70.1%	122,298	61.7%	125,791	64.5%	404,901	68.9%	183,259	72.1%
Regular day admission	13,709	5.9%	21,794	11.0%	24,413	12.5%	9,905	1.7%	10,897	4.3%
Regular night admission	15	0.0%	1	0.0%	1,480	0.8%	10	0.0%	12	0.0%
UNKNOWN	3	0.0%	1	0.0%	14	0.0%	0	0.0%	16	0.0%
Total	**230,492**		**198,218**		**194,988**		**587,737**		**254,027**	

Primary_Diagnosis

All sources supplied, at minimum, a primary diagnosis coded within the ICD-10 coding system. The ICD-10 codes with a trailing "X" or "-" were truncated to the most significant digit of diagnosis available, as a step in the cleaning process. Most PCT data sets included secondary diagnosis, and also included multiple records assigning co-morbidities to the primary diagnosis. As a unit of analysis, ICD-10 groupings were assigned, based on 2- and 3-digit numerical depth. For a detailed listing of diagnosis codes, please see eMedia\Dictionary\IP\ICD-10_diagnosis.csv.

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
Valid	210,147	91.2%	178,791	90.2%	181,051	92.9%	554332	94.3%	243,298	95.8%
Missing	8,850	3.8%	6,059	3.1%	3,717	1.9%	912	0.2%	10,688	4.2%
Invalid	11,495	5.0%	13,368	6.7%	10,220	5.2%	32,493	5.5%	41	0.0%
Total	**230,492**		**198,218**		**194,988**		**587,737**		**254,027**	

Secondary_Diagnosis [1-5]

Although PCT1 supplied only primary ICD-10 diagnosis, all other PCTs included secondary diagnosis with multiple records where necessary to specify multiple co-morbidities. For modelling purposes, secondary diagnosis was treated as identical to primary diagnosis for risk assessment. For a detailed listing of diagnosis codes, please see eMedia\Dictionary\IP\ICD-10_diagnosis.csv.

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
Sec DX1	-	-	86,904	43.9%	48,903	25.1%	244,448	41.6%	119,682	47.1%
Missing	230,492	100.0%	100,318	50.6%	139,109	71.3%	313,152	53.3%	123,267	48.5%
Unknown	-	-	10,996	5.5%	6,976	3.6%	30,137	5.1%	11,078	4.4%
Total	**230,492**		**198,218**		**194,988**		**587,737**		**254,027**	
Sec DX2	-	-	38,323	19.3%	68,212	34.9%	134,285	22.9%	76,564	30.1%
Missing	230,492	100.0%	156,173	78.8%	108,151	55.5%	444,347	75.6%	175,513	69.1%
Unknown	-	-	3,722	1.9%	18,625	9.6%	9,105	1.5%	1,950	0.8%
Total	**230,492**		**198,218**		**194,988**		**587,737**		**254,027**	
Sec DX3	-	-	21,778	11.0%	39,466	20.3%	69,407	11.8%	42,785	16.8%
Missing	230,492	100.0%	174,029	87.8%	151,577	77.7%	513,279	87.3%	210,284	82.8%
Unknown	-	-	2,411	1.2%	3,945	2.0%	5,051	0.9%	958	0.4%
Total	**230,492**		**198,218**		**194,988**		**587,737**		**254,027**	

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
Sec DX4	-	-	12,326	6.2%	20,379	10.5%	34,703	5.9%	24,253	9.5%
Missing	230,492	100.0%	184,781	93.2%	172,023	88.2%	550,704	93.7%	229,386	90.3%
Unknown	-	-	1,111	0.6%	2,586	1.3%	2,330	0.4%	388	0.2%
Total	**230,492**		**198,218**		**194,988**		**587,737**		**254,027**	
Sec DX5	-	-	8,676	4.4%	12,792	6.5%	16,288	2.8%	12,892	5.1%
Missing	230,492	100.0%	188,522	95.1%	180,859	92.8%	570,292	97.0%	240,892	94.8%
Unknown	-	-	1,020	0.5%	1,337	0.7%	1,157	0.2%	243	0.1%
Total	**230,492**		**198,218**		**194,988**		**587,737**		**254,027**	

Sex

A notation of "0" signifies that the sex of a patient has not been recorded, and "9" indicates that it is indeterminate, i.e., unable to be classified as either male or female.

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
Female	133,944	58.1%	104,338	52.6%	108,624	55.7%	334,354	56.9%	141,702	55.8%
Male	96,487	41.9%	93,879	47.4%	86,342	44.3%	253,372	43.1%	112,318	44.2%
Missing	44	0.0%	0	0.0%	0	0.0%	1	0.0%	0	0.0%
Not Known	0	0.0%	0	0.0%	9	0.0%	10	0.0%	5	0.0%
Not Specified	16	0.0%	1	0.0%	13	0.0%	0	0.0%	0	0.0%
Invalid	1	0.0%	0	0.0%	0	0.0%	0	0.0%	2	0.0%
Total	**230,492**		**198,218**		**194,988**		**587,737**		**254,027**	

TRANSFORMATION – IP DATA

ICD-10 diagnosis was modified to remove trailing “X” and “-“. The ICD-10 diagnoses (primary and secondary) that could not be validated were removed from consideration, although records of the admission were maintained. NHS numbers corresponding to an encrypted blank were removed from consideration. Verification of the encrypted value to be removed was obtained by encrypting a blank value with one-way SHA1 encryption by the PCT supplying the initial data, and reporting the encrypted value.

Duplicate records with identical information were removed.

OUTPATIENT (OP) DATA

LAYOUT – OP DATA

In order to ensure a consistent framework for modelling, data were generated according to a common and consistent format, as outlined below. A ‘Y’ under the column ‘Required for Model’ indicates variables which are needed to generate the model parameters.

Field Name	Definition	Required for Model?
Admin_Category	Type of patient (NHS, Private, etc.)	
Attendance_ID	Unique identifier for visit	
Attended_or_DNAd	This indicates whether or not a PATIENT attended for an appointment. If the PATIENT did not attend it also indicates whether or not advanced warning was given.	Y
Specialty_Code	Specialty of the consultant	Y
Date_Appointment_Request	Date appointment was requested	
Date_of_Attendance	Date patient attended	Y
Date_of_Birth	Date of patient birth	
Date_of_Death	Date of patient death	

Field Name	Definition	Required for Model?
Detention_Category	Mental category (mental illness, mental impairment, etc.)	
Diag_Version	This is used in the Clinical Information Group of the CDS to denote the scheme basis of a Diagnosis	
District_of_Residence_Code		
Ethnic_Origin	The new national code must be entered as the first character in the 2-character field. The second character is an optional field only required for use locally. It must, however, be able to be grouped consistently with the 16 main categories.	
First_Attendance	This indicates whether a patient is making a first or follow-up attendance	
Grade_of_Staff_Seeing_Pat	The grade of the health care professional seeing the patient	
Last_DNA_Date	Last appointment date that the patient did not attend	
Legal_Status	The classification is required for all patients who have a Hospital Provider Spell that includes the care of a CONSULTANT in the psychiatric specialties or have been discharged from such a Hospital Provider Spell and are required to receive supervised aftercare under the provisions of the Mental Health (Patients in the Community) Act 1995	
Location_Type	A code identifying the type of LOCATION	
NHS_Number	NHS number if known	Y
Outcome_of_Attendance	This records the outcome of an Out-Patient Attendance Consultant	
Postcode_of_Usual_Address	The code allocated by the Post Office to identify a group of postal delivery points	
Practice_Code	The GP practice where the patient is registered	
Primary_Diagnosis	The main condition treated or investigated during the relevant episode of health care where there is no definitive diagnosis, i.e., the main symptom, abnormal findings or problem (ICD-10)	
Primary_Procedure_Code	OPCS4 code for the primary procedure	
Priority_Type	This is the priority of a request for services	
Procedure_Status	OPERATION STATUS (procedure status) should be used once for each record to record states of knowledge regarding the operative procedure	
Procedure_Version	Procedure scheme in use: This is used in the Clinical Activity Group of the CDS to denote the scheme basis of an Intervention, Operation or A&E Treatment	
Provider_ID	Facility providing the services	
Purchaser_ID	Usually a PCT identifier	
Reason_for_Referral_Code	Also known as service type requested	
Referrer_Code	The ID of the health care professional making the referral	
Sex	Sex of patient	Y
Source_of_Referral_Code	A classification which is used to identify the source of referral of each Consultant Out-Patient Episode	Y
Subsidiary_Diagnosis	Secondary diagnosis (ICD-10 coding format)	

ASSESSMENT – OP DATA

Each data element likely to be involved in the modelling process was examined, and distributions evaluated. (For a detailed listing of codes and their associated meanings, please refer to the eight .csv files contained in the eMedia\Dictionary\OP folder.) When constraining data from each PCT to include only those records for the interval of 1 April 2002 to 31 March 2005, the following counts were observed:

PCT	Received	Within Interval	Duplicates	Remaining
PCT1	1,704,441	1,704,441	362,460	1,341,981
PCT2	798,180	595,169	11,809	583,360
PCT3	580,874	580,874	2,235	578,639
PCT4	1,326,298	1,312,624	15,569	1,297,055
PCT5	768,521	756,382	784	755,598
Total	**5,178,314**	**4,949,490**	**392,857**	**4,556,633**

Admin_Category

The admin_category data element distinguishes the type of patient being seen by outpatient services.

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
Amenity patient	5	0.0%	73	0.0%	0	0.0%	134	0.0%	0	0.0%
Category II patient	1	0.0%	0	0.0%	0	0.0%	3	0.0%	0	0.0%
Guy's & St Thomas's	3,159	0.2%	621	0.1%	0	0.0%	2,567	0.2%	0	0.0%
Missing	140	0.0%	10	0.0%	578,639	100.0%	9,544	0.7%	755,598	100.0%
NHS patient	1,305,813	97.3%	580,856	99.6%	0	0.0%	1,180,294	91.0%	0	0.0%
Not applicable	0	0.0%	323	0.1%	0	0.0%	0	0.0%	0	0.0%
Not known	11	0.0%	8	0.0%	0	0.0%	89,849	6.9%	0	0.0%
Papworth NHS pt	27,806	2.1%	94	0.0%	0	0.0%	12,008	0.9%	0	0.0%
Papworth Private pt	241	0.0%	0	0.0%	0	0.0%	5	0.0%	0	0.0%
Private patient	4,780	0.4%	1,371	0.2%	0	0.0%	1,938	0.1%	0	0.0%
Second episode	1	0.0%	0	0.0%	0	0.0%	262	0.0%	0	0.0%
UNKNOWN	24	0.0%	4	0.0%	0	0.0%	451	0.0%	0	0.0%
Total	**1,341,981**		**583,360**		**578,639**		**1,297,055**		**755,598**	

Attended_or_DNAd

This data element indicates whether or not a patient attended for an appointment. If the patient did not attend it also indicates whether or not advanced warning was given.

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
Appointment cancelled (by patient)	132,189	9.9%	58,720	10.1%	0	0.0%	5,779	0.4%	61,558	8.1%
Appointment cancelled (by provider)	157,993	11.8%	88,187	15.1%	0	0.0%	6,069	0.5%	63,886	8.5%
Arrived late	4,188	0.3%	1,439	0.2%	3,637	0.6%	1,424	0.1%	26,781	3.5%
Attended on time	904,575	67.4%	406,024	69.6%	575,002	99.4%	1,198,902	92.4%	470,637	62.3%
Did not attend – no advance warning given	137,133	10.2%	28,076	4.8%	0	0.0%	83,720	6.5%	119,116	15.8%
Patient arrived late (not seen)	680	0.1%	268	0.0%	0	0.0%	51	0.0%	1,391	0.2%
Unknown	5,223	0.4%	646	0.1%	0	0.0%	1,110	0.1%	12,229	1.6%
Total	**1,341,981**		**583,360**		**578,639**		**1,297,055**		**755,598**	

Specialty_Code

This is the specialty in which the consultant is contracted or recognised, and classifies clinical work divisions more precisely for a limited number of specialties.

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
Accident & Emergency	629	0.0%	3,477	0.6%	0	0.0%	5,754	0.4%	57	0.0%
Anaesthetics	3,494	0.3%	920	0.2%	3,693	0.6%	5,634	0.4%	2,364	0.3%
Ante-Natal Obstetrics	684	0.1%	16,479	2.8%	3,576	0.6%	7,696	0.6%	20,423	2.7%
Audiological Medicine	2,133	0.2%	19	0.0%	1,379	0.2%	2,455	0.2%	636	0.1%
Cardiology	54,005	4.0%	19,870	3.4%	16,182	2.8%	20,278	1.6%	17,677	2.3%
Cardiothoracic Surgery	2,407	0.2%	1,722	0.3%	1,911	0.3%	3,907	0.3%	1,271	0.2%
Chemical Pathology	4,064	0.3%	120	0.0%	0	0.0%	301	0.0%	0	0.0%
Child & Adolescent Psych	146	0.0%	24	0.0%	0	0.0%	5,760	0.4%	70	0.0%
Clinical Genetics	307	0.0%	475	0.1%	0	0.0%	404	0.0%	31	0.0%
Clinical Immunology	174	0.0%	297	0.1%	0	0.0%	884	0.1%	501	0.1%
Clinical Neuro-Physiology	1,354	0.1%	23	0.0%	457	0.1%	4,950	0.4%	1,676	0.2%
Clinical Pharmacology	243	0.0%	33	0.0%	0	0.0%	19	0.0%	41	0.0%
Clinical Physiology	1	0.0%	0	0.0%	18	0.0%	0	0.0%	9	0.0%
Community Medicine	6	0.0%	896	0.2%	0	0.0%	59	0.0%	18	0.0%
Dental Medicine	618	0.0%	13	0.0%	1,040	0.2%	765	0.1%	1,833	0.2%
Dermatology	66,446	5.0%	36,668	6.3%	37,450	6.5%	70,639	5.4%	52,873	7.0%
Endocrinology	17,421	1.3%	16,770	2.9%	29,570	5.1%	1,422	0.1%	20,115	2.7%
ENT	62,438	4.7%	33,165	5.7%	20,045	3.5%	63,338	4.9%	49,535	6.6%
Forensic Psychiatry	89	0.0%	0	0.0%	0	0.0%	8	0.0%	0	0.0%
Gastroenterology	40,939	3.1%	14,578	2.5%	20,464	3.5%	5,212	0.4%	27,121	3.6%
General Medicine	79,076	5.9%	9,003	1.5%	18,188	3.1%	144,645	11.2%	22,975	3.0%
General Practice	0	0.0%	0	0.0%	-	0.0%	142	0.0%	0	0.0%
General Surgery	139,751	10.4%	47,369	8.1%	34,182	5.9%	103,531	8.0%	38,651	5.1%
Genito-urinary Medicine	144	0.0%	4	0.0%	0	0.0%	19	0.0%	32	0.0%
Geriatric Medicine	7,741	0.6%	8,112	1.4%	8,608	1.5%	598	0.0%	3,069	0.4%
Gynaecology	67,218	5.0%	21,618	3.7%	22,257	3.8%	50,391	3.9%	50,881	6.7%
Haematology	6,582	0.5%	3	0.0%	5,076	0.9%	49	0.0%	69	0.0%
Haematology (Clinical)	24,506	1.8%	13,600	2.3%	12,703	2.2%	38,066	2.9%	53,364	7.1%
Histopathology	2	0.0%	0	0.0%	0	0.0%	2	0.0%	1	0.0%
Immunopathology	2,697	0.2%	0	0.0%	581	0.1%	208	0.0%	5	0.0%
Infectious Diseases	587	0.0%	30	0.0%	0	0.0%	377	0.0%	506	0.1%
Maternity Function	0	0.0%	1	0.0%	0	0.0%	32	0.0%	103	0.0%
Medical Microbiology	37	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
Medical Oncology	16,488	1.2%	735	0.1%	4,484	0.8%	4,773	0.4%	3,566	0.5%

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
Mental Handicap	136	0.0%	923	0.2%	0	0.0%	3,075	0.2%	7	0.0%
Mental Illness	1,541	0.1%	63	0.0%	6	0.0%	32,027	2.5%	147	0.0%
Midwifery	312	0.0%	5	0.0%	6,297	1.1%	8,384	0.6%	434	0.1%
Nephrology	24,379	1.8%	5,002	0.9%	10,957	1.9%	15,491	1.2%	8,140	1.1%
Neurology	21,758	1.6%	15,247	2.6%	12,133	2.1%	17,346	1.3%	15,275	2.0%
Neurosurgery	2,349	0.2%	1,111	0.2%	1,051	0.2%	9,383	0.7%	3,485	0.5%
Nuclear Medicine	329	0.0%	1	0.0%	2,060	0.4%	3	0.0%	28	0.0%
Obstetrics & Gynaecology	995	0.1%	4	0.0%	7,400	1.3%	72	0.0%	168	0.0%
Obstetrics For Bed Or Del	50,560	3.8%	7,931	1.4%	16,944	2.9%	40,078	3.1%	15,940	2.1%
Occupational Medicine	15	0.0%	0	0.0%	0	0.0%	0	0.0%	1	0.0%
Ophthalmology	153,390	11.4%	81,044	13.9%	42,247	7.3%	148,938	11.5%	54,435	7.2%
Oral Surgery	24,440	1.8%	13,741	2.4%	11,476	2.0%	38,629	3.0%	8,853	1.2%
Orthodontics	19,297	1.4%	6,815	1.2%	8,438	1.5%	31,017	2.4%	11,093	1.5%
Other (than Maternity)	0	0.0%	0	0.0%	0	0.0%	161	0.0%	0	0.0%
Paediatric Dentistry	1,868	0.1%	6	0.0%	4,587	0.8%	1,671	0.1%	6,188	0.8%
Paediatric Neurology	907	0.1%	58	0.0%	1,899	0.3%	838	0.1%	4,190	0.6%
Paediatric Surgery	1,563	0.1%	4,810	0.8%	2,457	0.4%	1,934	0.1%	4,861	0.6%
Paediatrics	39,396	2.9%	20,342	3.5%	14,817	2.6%	40,732	3.1%	29,351	3.9%
Pain Management	3,054	0.2%	5,711	1.0%	1,517	0.3%	5,254	0.4%	2,411	0.3%
Palliative Medicine	160	0.0%	413	0.1%	127	0.0%	95	0.0%	1	0.0%
Plastic Surgery	16,752	1.2%	17,518	3.0%	4,612	0.8%	32,793	2.5%	9,958	1.3%
Post-Natal Obstetrics	0	0.0%	182	0.0%	46	0.0%	7	0.0%	228	0.0%
Psychogeriatrics (ESMI)	151	0.0%	0	0.0%	0	0.0%	4,661	0.4%	0	0.0%
Psychotherapy	46	0.0%	11	0.0%	0	0.0%	3,762	0.3%	16	0.0%
Radiology	1,154	0.1%	44	0.0%	0	0.0%	1,022	0.1%	180	0.0%
Radiotherapy	19,048	1.4%	12,739	2.2%	4,843	0.8%	33,491	2.6%	17,139	2.3%
Rehabilitation	2,679	0.2%	460	0.1%	5,421	0.9%	3,591	0.3%	165	0.0%
Restorative Dentistry	27,989	2.1%	529	0.1%	12,949	2.2%	7,915	0.6%	10,082	1.3%
Rheumatology	33,971	2.5%	23,602	4.0%	21,908	3.8%	25,120	1.9%	34,844	4.6%
Thoracic Medicine	14,693	1.1%	15,460	2.7%	14,987	2.6%	3,638	0.3%	27,330	3.6%
Trauma & Orthopaedics	120,014	8.9%	66,472	11.4%	66,495	11.5%	187,655	14.5%	60,775	8.0%
Urology	60,863	4.5%	26,794	4.6%	20,034	3.5%	36,746	2.8%	25,986	3.4%
UNKNOWN	95,365	7.1%	10,296	1.8%	41,067	7.1%	19,107	1.5%	33,600	4.4%
Missing	380	0.0%	2	0.0%	0	0.0%	101	0.0%	814	0.1%
Total	**1,341,981**		**583,360**		**578,639**		**1,297,055**		**755,598**	

First_Attendance

This indicates whether a patient is making a first attendance or contact; or a follow-up attendance or contact.

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
First attendance	401,641	29.9%	167,161	28.7%	165,555	28.6%	395,436	30.5%	234,753	31.1%
Follow-up attendance	935,071	69.7%	415,858	71.3%	413,084	71.4%	900,966	69.5%	517,936	68.5%
Unknown	5,269	0.4%	341	0.1%	0	0.0%	653	0.1%	2,909	0.4%
Total	**1,341,981**		**583,360**		**578,639**		**1,297,055**		**755,598**	

Outcome_of_Attendance

This records the outcome of an outpatient attendance consultant.

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
Another appointment given	429,629	32.0%	40,632	7.0%	0	0.0%	625,146	48.2%	316,796	41.9%
Appointment to be made at a later date	377,230	28.1%	453,458	77.7%	0	0.0%	247,070	19.0%	159,377	21.1%
Discharged from consultant's care (last attendance)	159,043	11.9%	83,024	14.2%	0	0.0%	335,218	25.8%	160,106	21.2%
Missing	362,182	27.0%	0	0.0%	578,639	100.0%	83,386	6.4%	0	0.0%
UNKNOWN	13,897	1.0%	6,246	1.1%	0	0.0%	6,235	0.5%	119,319	15.8%
Total	**1,341,981**		**583,360**		**578,639**		**1,297,055**		**755,598**	

Priority_Type

This is the priority of a request for services; in the case of services to be provided by a consultant, it is as assessed by or on behalf of the consultant.

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
Missing	62,633	4.7%	0	0.0%	578,639	100.0%	18,436	1.4%	0	0.0%
Routine	852,927	63.6%	507,633	87.0%	0	0.0%	1,125,907	86.8%	595,333	78.8%
UNKNOWN	276	0.0%	13,867	2.4%	0	0.0%	731	0.1%	6,155	0.8%
Urgent	426,145	31.8%	61,860	10.6%	0	0.0%	151,981	11.7%	154,110	20.4%
Total	**1,341,981**		**583,360**		**578,639**		**1,297,055**		**755,598**	

TRANSFORMATION – OP DATA

ICD-10 diagnosis was modified to remove trailing “X” and “-“. The ICD-10 diagnoses (primary and secondary) that could not be validated were removed from consideration, although records of the outpatient attendance were maintained.

NHS numbers corresponding to the encrypted blank value were removed from consideration. Verification of the encrypted value to be removed was obtained by encrypting a blank value with one-way SHA1 encryption by the PCT supplying the initial data, and reporting the encrypted value.

Duplicate records with identical information were removed. In the OP dataset, duplicates were encountered for the year of 2003, due to duplication of a dataset in submission with an alternate form of dates. Duplicates were identified and removed based on comparison of unique identifiers. Year of birth, which was originally used as a part of a unique identifier, was encrypted from two different formats, resulting in two different values. However, when this was excluded as a comparator, approximately 328,000 duplicate records were removed. (The use of duplicate information would result in incorrect identification of variables during the data mining phase of the modelling process.)

ACCIDENT & EMERGENCY (A&E) DATA

LAYOUT – A&E DATA

In order to ensure a consistent framework for modelling, data were generated according to a common and consistent format, as outlined below. A 'Y' under the column 'Required for Model' indicates variables which are needed to generate the model parameters.

Field Name	Definition	Required for Model?
Att_ArrivalMode_NatCode	Transportation to AE (ambulance or other)	Y
Att_Category_NatCode	First incident, planned follow-up or unplanned follow-up	
Att_Date_Arrival	Date of event	Y
Att_Diag_01_NatCode	National diagnostic codes	Y
Att_Diag_01_Side_NatCode	National diagnostic codes	
Att_Diag_01_Site_NatCode	National diagnostic codes	
Att_Diag_02_NatCode	National diagnostic codes	Y
Att_Diag_02_Side_NatCode	National diagnostic codes	
Att_Diag_02_Site_NatCode	National diagnostic codes	
Att_Disposal_NatCode	A coding of the ways in which an ACCIDENT AND EMERGENCY ATTENDANCE might end	Y
Att_IncidentLocation_NatCode	Where the incident occurred (home, work, etc.)	
Att_Initiator_NatCode		
Att_Investigation_01_NatCode	Test performed, (X-ray, ultrasound, etc.)	Y
Att_Investigation_02_NatCode	Test performed, (X-ray, ultrasound, etc.)	Y
Att_Investigation_03_NatCode	Test performed, (X-ray, ultrasound, etc.)	Y
Att_Investigation_04_NatCode	Test performed, (X-ray, ultrasound, etc.)	Y
Att_Investigation_05_NatCode	Test performed, (X-ray, ultrasound, etc.)	Y
Att_Investigation_06_NatCode	Test performed, (X-ray, ultrasound, etc.)	Y
Att_PatientGroup_NatCode	Grouping of type of incident (road traffic, sports, etc.)	
Att_Treatment_01_NatCode	Treatment - (prescription, splint, etc.)	
Att_Treatment_02_NatCode	Treatment - (prescription, splint, etc.)	
Att_Type		
In_AE_over_4_hours	Logical field	
Pat_Age	Age in years	Y
Pat_NHSNo	NHS number if known	Y
Pat_Sex_NatCode	Sex of patient	Y
Reattendance_within_1_month	Logical field	
Reattendance_within_12_months	Logical field	
Reattendance_within_6_months	Logical field	

ASSESSMENT – A&E DATA

A significant number of records had no NHS number associated. These records had to be removed, as defined below: (Please note that all frequencies reflect the removal of blank records).

Each data element likely to be involved in the modelling process was examined, and distributions evaluated. (For a detailed listing of codes and their associated meanings, please refer to the six .csv files contained in the eMedia\Dictionary\AE folder.) Constraining data from each PCT to include only those records for the interval from 1 April 2002 through 31 March 2005, the following counts were observed:

PCT	Received	Within Interval	Duplicates	Remaining
PCT1	316,662	255,284	143	255,141
PCT2	2,125	2,125	0	2,125
PCT3	-	-	-	-
PCT4	133,924	119,174	18	119,156
PCT5	180,965	180,572	149	180,423
Total	**633,676**	**618,533**	**310**	**556,845**

Att_ArrivalMode_NatCode

This variable identifies the means by which a patient arrived at an Accident and Emergency department.

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
Brought in by ambulance	66,044	25.9%	514	24.2%	-	-	27,817	23.3%	32,887	18.2%
Other	189,095	74.1%	1,611	75.8%	-	-	91,338	76.7%	145,632	80.7%
Unknown	2	0.0%	0	0.0%	-	-	1	0.0%	1,904	1.1%
Total	**255,141**		**2,125**		**-**		**119,156**		**180,423**	

Att_Category_NatCode

An indication of whether a patient is making a first or follow-up attendance at a particular Accident and Emergency department.

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
First A&E Attendance	243,980	95.6%	2,117	99.6%	-	-	113,207	95.0%	178,513	98.9%
Follow-up A&E Attendance - planned	4,750	1.9%	1	0.0%	-	-	2,360	2.0%	1,016	0.6%
Follow-up A&E Attendance - unplanned	6,410	2.5%	7	0.3%	-	-	3,588	3.0%	893	0.5%
Unknown	1	0.0%	0	0.0%	-	-	1	0.0%	1	0.0%
Total	**255,141**		**2,125**		**-**		**119,156**		**180,423**	

Att_Diag_01_NatCode

The first recorded patient diagnosis for an Accident and Emergency attendance. This is required for recording within an Accident and Emergency attendance CDS. (Please note that valid entries for this variable are noted in eMedia\Dictionary\AE\AE_Diag.csv.)

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
Allergy (including anaphylaxis)	1,410	0.6%	16	0.8%	-	-	1,002	0.8%	2	0.0%
Bites/stings	1,575	0.6%	16	0.8%	-	-	877	0.7%	214	0.1%
Burns and scalds	2,009	0.8%	11	0.5%	-	-	1,418	1.2%	33,023	18.3%
Cardiac conditions	6,721	2.6%	87	4.1%	-	-	3,476	2.9%	145	0.1%
Central nervous system conditions (excluding strokes)	2,909	1.1%	15	0.7%	-	-	1,244	1.0%	2	0.0%

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
Cerebro-vascular conditions	1,840	0.7%	29	1.4%	-	-	928	0.8%	685	0.4%
Contusion/abrasion	12,595	4.9%	142	6.7%	-	-	3,758	3.2%	5	0.0%
Dermatological conditions	1,796	0.7%	1	0.0%	-	-	292	0.2%	109	0.1%
Diabetes and other endocrinological conditions	669	0.3%	6	0.3%	-	-	207	0.2%	1,630	0.9%
Diagnosis not classifiable	21,288	8.3%	395	18.6%	-	-	13,870	11.6%	16	0.0%
Dislocation/fracture/joint injury/amputation	15,422	6.0%	377	17.7%	-	-	14,301	12.0%	17	0.0%
ENT conditions	2,360	0.9%	46	2.2%	-	-	1,411	1.2%	2	0.0%
Electric shock	56	0.0%	1	0.0%	-	-	20	0.0%	3,752	2.1%
Facio-maxillary conditions	574	0.2%	1	0.0%	-	-	242	0.2%	34	0.0%
Foreign body	3,208	1.3%	49	2.3%	-	-	2,608	2.2%	0	0.0%
Gastrointestinal conditions	9,659	3.8%	72	3.4%	-	-	3,975	3.3%	4	0.0%
Gynaecological conditions	3,346	1.3%	10	0.5%	-	-	603	0.5%	3	0.0%
Haematological conditions	473	0.2%	2	0.1%	-	-	413	0.3%	2	0.0%
Head injury	5,711	2.2%	106	5.0%	-	-	6,213	5.2%	8	0.0%
Infectious disease	2,306	0.9%	8	0.4%	-	-	431	0.4%	1	0.0%
Laceration	19,488	7.6%	177	8.3%	-	-	11,189	9.4%	8	0.0%
Local infection	7,597	3.0%	9	0.4%	-	-	5,763	4.8%	6	0.0%
Missing	82,600	32.4%	35	1.6%	-	-	9,702	8.1%	44	0.0%
Muscle/tendon injury	2,407	0.9%	51	2.4%	-	-	4,112	3.5%	0	0.0%
Near drowning	7	0.0%	1	0.0%	-	-	4	0.0%	0	0.0%
Nerve injury	135	0.1%	3	0.1%	-	-	117	0.1%	0	0.0%
Nothing abnormal detected	3,526	1.4%	27	1.3%	-	-	2,100	1.8%	4	0.0%
Obstetric conditions	934	0.4%	7	0.3%	-	-	218	0.2%	0	0.0%
Ophthalmological conditions	1,482	0.6%	56	2.6%	-	-	3,105	2.6%	3	0.0%
Other vascular conditions	929	0.4%	6	0.3%	-	-	369	0.3%	0	0.0%
Poisoning (including overdose)	1,818	0.7%	43	2.0%	-	-	1,978	1.7%	0	0.0%
Psychiatric conditions	1,325	0.5%	3	0.1%	-	-	486	0.4%	1	0.0%
Respiratory conditions	10,940	4.3%	55	2.6%	-	-	2,492	2.1%	11	0.0%
Septicaemia	204	0.1%	0	0.0%	-	-	50	0.0%	0	0.0%
Social problem (includes chronic alcoholism and homelessness)	1,144	0.4%	1	0.0%	-	-	156	0.1%	1	0.0%
Soft tissue inflammation	7,318	2.9%	42	2.0%	-	-	9,148	7.7%	7	0.0%
Sprain/ligament injury	13,524	5.3%	166	7.8%	-	-	9,103	7.6%	12	0.0%
UNKNOWN	0	0.0%	24	1.1%	-	-	0	0.0%	140,667	78.0%
Urological conditions (including cystitis)	3,669	1.4%	29	1.4%	-	-	1,591	1.3%	3	0.0%
Vascular injury	112	0.0%	0	0.0%	-	-	166	0.1%	0	0.0%
Visceral injury	55	0.0%	0	0.0%	-	-	18	0.0%	2	0.0%
Total	**255,141**		**2,125**		**-**		**119,156**		**180,423**	

ATT_DIAG_02_NATCODE

The second recorded patient diagnosis for an Accident and Emergency attendance. This is required for recording within an Accident and Emergency attendance CDS. (Please note that valid entries for this variable are noted in eMedia\Dictionary\AE\AE_Diag.csv.)

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
Allergy (including anaphylaxis)	21	0.0%	0	0.0%	-	-	35	0.0%	2	0.0%
Bites/stings	98	0.0%	0	0.0%	-	-	28	0.0%	214	0.1%
Burns and scalds	165	0.1%	2	0.1%	-	-	62	0.1%	33,023	18.3%
Cardiac conditions	61	0.0%	2	0.1%	-	-	44	0.0%	145	0.1%
Central Nervous System conditions (excluding strokes)	35	0.0%	0	0.0%	-	-	28	0.0%	2	0.0%
Cerebro-vascular conditions	39	0.0%	0	0.0%	-	-	18	0.0%	685	0.4%
Contusion/abrasion	534	0.2%	11	0.5%	-	-	275	0.2%	5	0.0%
Dermatological conditions	85	0.0%	0	0.0%	-	-	4	0.0%	109	0.1%
Diabetes and other endocrinological conditions	13	0.0%	0	0.0%	-	-	3	0.0%	1,630	0.9%
Diagnosis not classifiable	497	0.2%	16	0.8%	-	-	267	0.2%	16	0.0%
Dislocation/fracture/joint injury/amputation	357	0.1%	16	0.8%	-	-	395	0.3%	17	0.0%
ENT conditions	49	0.0%	0	0.0%	-	-	27	0.0%	2	0.0%
Electric shock	4	0.0%	0	0.0%	-	-	1	0.0%	3,752	2.1%
Facio-maxillary conditions	21	0.0%	0	0.0%	-	-	11	0.0%	34	0.0%
Foreign body	56	0.0%	0	0.0%	-	-	51	0.0%	0	0.0%
Gastrointestinal conditions	123	0.0%	2	0.1%	-	-	57	0.0%	4	0.0%
Gynaecological conditions	32	0.0%	0	0.0%	-	-	10	0.0%	3	0.0%
Haematological conditions	13	0.0%	0	0.0%	-	-	6	0.0%	2	0.0%
Head injury	751	0.3%	8	0.4%	-	-	445	0.4%	8	0.0%
Infectious disease	10	0.0%	0	0.0%	-	-	4	0.0%	1	0.0%
Laceration	290	0.1%	12	0.6%	-	-	657	0.6%	8	0.0%
Local infection	171	0.1%	0	0.0%	-	-	139	0.1%	6	0.0%
Missing	250,613	98.2%	2,037	95.9%	-	-	115,588	97.0%	44	0.0%
Muscle/tendon injury	115	0.0%	3	0.1%	-	-	157	0.1%	0	0.0%
Near drowning	1	0.0%	0	0.0%	-	-	0	0.0%	0	0.0%
Nerve injury	16	0.0%	0	0.0%	-	-	13	0.0%	0	0.0%
Nothing abnormal detected	25	0.0%	1	0.0%	-	-	23	0.0%	4	0.0%
Obstetric conditions	14	0.0%	0	0.0%	-	-	5	0.0%	0	0.0%
Ophthalmological conditions	48	0.0%	2	0.1%	-	-	41	0.0%	3	0.0%
Other vascular conditions	37	0.0%	1	0.0%	-	-	6	0.0%	0	0.0%
Poisoning (including overdose)	16	0.0%	1	0.0%	-	-	57	0.0%	0	0.0%
Psychiatric conditions	22	0.0%	0	0.0%	-	-	24	0.0%	1	0.0%
Respiratory conditions	235	0.1%	1	0.0%	-	-	45	0.0%	11	0.0%
Septicaemia	4	0.0%	0	0.0%	-	-	1	0.0%	0	0.0%
Social problem (includes chronic alcoholism and homelessness)	27	0.0%	0	0.0%	-	-	4	0.0%	1	0.0%
Soft tissue inflammation	199	0.1%	0	0.0%	-	-	346	0.3%	7	0.0%
Sprain/ligament injury	258	0.1%	5	0.2%	-	-	242	0.2%	12	0.0%
UNKNOWN	0	0.0%	3	0.1%	-	-	0	0.0%	140,667	78.0%
Urological conditions (including cystitis)	82	0.0%	2	0.1%	-	-	32	0.0%	3	0.0%
Vascular injury	3	0.0%	0	0.0%	-	-	4	0.0%	0	0.0%
Visceral injury	1	0.0%	0	0.0%	-	-	1	0.0%	2	0.0%
Total	255,141		2,125		-		119,156		180,423	

Att_Investigation_[1-6]_NatCode

The first recorded Accident and Emergency investigation of a clinical intervention for an Accident and Emergency attendance. This is required for recording within an Accident and Emergency attendance CDS.

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
Bacteriology	0	0.0%	2	0.1%	-	-	534	0.4%	1	0.0%
Biochemistry	18,963	6.1%	9	0.4%	-	-	5,003	3.8%	9	0.0%
Computerised Tomography	169	0.1%	2	0.1%	-	-	30	0.0%	0	0.0%
Cross match	775	0.2%	1	0.0%	-	-	292	0.2%	0	0.0%
ECG	23,884	7.6%	8	0.4%	-	-	8,181	6.3%	13	0.0%
Haematology	23,848	7.6%	5	0.2%	-	-	5,313	4.1%	19	0.0%
Histology Total	0	0.0%	0	0.0%	-	-	26	0.0%	0	0.0%
Magnetic Resonance Imaging (MRI)	0	0.0%	0	0.0%	-	-	8,262	6.3%	4	0.0%
Other	40,978	13.1%	64	2.9%	-	-	0	0.0%	11	0.0%
Ultrasound	103	0.0%	0	0.0%	-	-	48	0.0%	71	0.0%
UNKNOWN	0	0.0%	24	1.1%	-	-	0	0.0%	19,551	10.6%
Urine	9,485	3.0%	0	0.0%	-	-	2,470	1.9%	4	0.0%
X-ray	63,764	20.4%	596	26.7%	-	-	40,972	31.3%	55	0.0%
Missing	131,144	41.9%	1,524	68.2%	-	-	59,711	45.6%	164,967	89.3%
Total	**313,113**		**2,235**		**-**		**130,842**		**184,705**	

Att_PatientGroup_NatCode

A coded classification to identify the reason for an Accident and Emergency episode.

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
Assault	5,405	2.1%	22	1.0%	-	-	2,164	1.8%	3,408	1.9%
Brought in dead	150	0.1%	0	0.0%	-	-	13	0.0%	1	0.0%
Deliberate self-harm	1,976	0.8%	34	1.6%	-	-	1,626	1.4%	1,193	0.7%
Firework injury	0	0.0%	0	0.0%	-	-	5	0.0%	11	0.0%
Other accident	83,283	32.6%	1,164	54.8%	-	-	41,062	34.5%	49	0.0%
Other than above	120,632	47.3%	659	31.0%	-	-	62,771	52.7%	171,890	95.3%
Road traffic accident	0	0.0%	82	3.9%	-	-	3,503	2.9%	1,697	0.9%
Sports injury	1,162	0.5%	164	7.7%	-	-	8,011	6.7%	2,150	1.2%
Unknown	42,533	16.7%	0	0.0%	-	-	1	0.0%	24	0.0%
Total	**255,141**		**2,125**		**-**		**119,156**		**180,423**	

In_AE_over_4_hours

This variable was calculated to determine whether or not a patient was more or less than four hours in the Accident and Emergency department before receiving treatment.

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
Less than 4 hours in AE	209,696	82.2%	1,922	90.4%	-	-	111,362	93.5%	152,364	84.4%
Missing	1	0.0%	0	0.0%	-	-	0	0.0%	0	0.0%
Over 4 hours in AE	45,444	17.8%	203	9.6%	-	-	7,794	6.5%	28,059	15.6%
Total	**255,141**		**2,125**		**-**		**119,156**		**180,423**	

Pat_Age

This variable represents the age of the patient.

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
0-14	60,035	23.5%	489	23.0%	-	-	14,019	11.8%	53,732	29.8%
15-44	107,077	42.0%	762	35.9%	-	-	35,078	29.4%	76,978	42.7%
45-64	41,885	16.4%	404	19.0%	-	-	15,963	13.4%	26,030	14.4%
65-74	17,604	6.9%	177	8.3%	-	-	5,999	5.0%	11,987	6.6%
75+	28,540	11.2%	293	13.8%	-	-	13,398	11.2%	11,696	6.5%
Invalid	0	0.0%	0	0.0%	-	-	5	0.0%	0	0.0%
Missing	0	0.0%	0	0.0%	-	-	34,694	29.1%	0	0.0%
Total	**255,141**		**2,125**		**-**		**119,156**		**180,423**	

Pat_Sex_NatCode

This variable is identical to the codes used for gender as defined for inpatient.

Category	PCT1	%	PCT2	%	PCT3	%	PCT4	%	PCT5	%
Female	127,069	49.8%	954	44.9%	-	-	54,217	45.5%	86,997	48.2%
Male	128,059	50.2%	1,171	55.1%	-	-	64,939	54.5%	93,416	51.8%
Unknown	13	0.0%	0	0.0%	-	-	0	0.0%	10	0.0%
Total	**255,141**		**2,125**		**-**		**119,156**		**180,423**	

TRANSFORMATION – A&E DATA

Only the first two bytes of the national code for diagnosis appeared valid. Site-specific coding appeared to be inconsistent and, at times, confounding. For example, site specific coding was sometimes combined with an illogical diagnosis, such as head lacerations on leg, cardiac condition on head, etc.

A large number of NHS numbers were unspecified, and blank values had to be removed from consideration.

GENERAL PRACTICE (GP) DATA

LAYOUT – GP DATA

In order to ensure a consistent framework for modelling, data were generated according to a common and consistent format, as outlined below. A 'Y' under the column 'Required for Model' indicates variables which are needed to generate the model parameters.

Field Name	Definition	Required for Model?
Active	Current registration status with the practice	**Y**
Age	Age at time of data extraction	**Y**
Date End	Date when the item ceased to apply	
Date of Birth	Patient date of birth	
Date Recorded	Date on which recorded	
Event Code	Clinical code (READ code) unless otherwise indicated	**Y**
Event Code Original	Presenting clinical code	
Event Date	Date to which the code applies	**Y**
Event ID	System assigned date specific to this encounter	
HCP	NHS specified ID of the health care professional responsible	
HCP Type	Profession of health care professional	
Internal ID	Unique record identifier internal to the practice system	

Field Name	Definition	Required for Model?
NHS Number	NHS number if known	**Y**
Patient Practice ID	Internal practice number for the patient	
Post Code	Post code of current address	
Practice ID	The ID of the practice	**Y**
Registered Date	Date patient registered with practice	
Removed Date	Date registration ended	
Sex	Sex of patient	**Y**
Value 1	First numeric value for Rx, amount prescribed	**Y**
Value 2	Second numeric value for Rx, daily dose	

ASSESSMENT – GP DATA

Each data element likely to be involved in the modelling process was examined, and distributions evaluated. (For a detailed listing of codes and their associated meanings, please refer to the six .csv files contained in the eMedia\Dictionary\GP folder.) Constraining data from each PCT to include only those records for the interval of 1 April 2002 to 31 March 2005, the following counts were observed:

PCT	Received	Within Interval	Duplicates	Remaining
PCT1	16,989,555	16,981,931	24,011	16,957,920
PCT4	22,897,392	8,523,293	63,570	8,459,723
Total	**39,886,947**	**25,505,224**	**87,581**	**25,417,643**

PCT2 was excluded from use as full permission could not be obtained for all data, resulting in a significant number of exclusions. This provided a sample size too small from PCT2 for effective statistical analysis. Because only PCT1 and PCT4 supplied GP data which could be used, other PCT data were excluded from this point onward.

Age

Where age was not provided, it was calculated by de-referencing other available data sources; and using a backward reference technique, where the encrypted birthdate was associated with a known age, and this age was applied to all similarly occurring records. For example, given a patient with an encrypted birthdate of “X”, and a known age of 10, then all patients with an equivalent encrypted birthdate of X were also 10 at the adjusted time of encounter. Utilising this method, approximately 97% of all birthdates were resolved successfully. It is assumed for unresolved cases that the birthdate might have been either not specified or specified out of a range suitable for other existing data.

Category	PCT1	%	PCT4	%
0-14	769,146	4.5%	2,642	0.0%
15-44	4,233,322	25.0%	10,370	0.1%
45-64	3,657,659	21.6%	37,834	0.4%
65-74	2,075,475	12.2%	7,768	0.1%
75+	2,123,398	12.5%	24,723	0.3%
Invalid	1,247,252	7.4%	8,376,386	99.0%
Missing	2,851,668	16.8%	0	0.0%
Total	**16,957,920**		**8,459,723**	

Event Code

Each record extracted from GP data has an associated event code, which is dependent on the version of the Read code system implemented at that site. In order to “normalise” the reference to a common base, each Read code was mapped to its equivalent in the Clinical Terms Version 3 (CTV3) Read terms, and grouped.

Level 1 - Categories	PCT1	%	PCT4	%
Occupations	17,086	0.1%	15,824	0.2%
Administration	1,230,405	7.3%	1,192,870	14.1%
Causes of injury and poisoning	9,103	0.1%	11,410	0.1%
Context-dependent categories	1,563,864	9.2%	1,223,914	14.5%
Attribute	6	0.0%	0	0.0%
Staging and scales	16,463	0.1%	4,519	0.1%
Additional values	131,536	0.8%	70,443	0.8%
Anatomical concepts	4,024	0.0%	2,583	0.0%
Organisms	2,038	0.0%	3,334	0.0%
Operations, procedures, and interventions	2,264,508	13.4%	1,980,925	23.4%
Extinct cross-type concept	71,140	0.4%	10,538	0.1%
Clinical findings	5,522,819	32.6%	3,941,897	46.6%
Appliances+equipment	228,822	1.3%	67	0.0%
Drug	5,478,503	32.3%	1,398	0.0%
Unknown / Invalid	417,603	2.5%	1	0.0%
Total	**16,957,920**		**8,459,723**	

Categorisation of the Read codes occurring at the second “tier” is represented below to show the distribution of types of data stored. In order to achieve results similar to the modelling performed, data should be similar. Notably, PCT4 did not supply drug data; drug data significantly help to further define the information available to evaluate a patient’s effective risk of admission.

Level 2 - Categories	PCT1	%	PCT4	%
History and observations	4,975,533	29.3%	3,204,721	37.9%
O/E - specified examination findings	1,150,975	6.8%	931,588	11.0%
Investigations	1,069,008	6.3%	987,320	11.7%
Regimes and therapies	928,072	5.5%	796,295	9.4%
Administrative statuses	891,911	5.3%	794,233	9.4%
Drug groups primarily affecting the cardiovascular system	1,631,373	9.6%	1	0.0%
Disorders	547,265	3.2%	737,134	8.7%
Hormones, synthetic substitutes, and antagonists	928,039	5.5%	3	0.0%
Administrative procedures	262,893	1.6%	322,844	3.8%
Anti-infectives	489,642	2.9%	5	0.0%
Drug groups primarily affecting the central nervous system	446,392	2.6%	10	0.0%
Drug groups primarily affecting the respiratory system	364,400	2.1%	1	0.0%
Drug groups primarily affecting the gastro-intestinal system	347,592	2.0%	2	0.0%
Analgesics and non-steroidal anti-inflammatory drugs	341,180	2.0%	0	0.0%
Operations and procedures	151,028	0.9%	138,305	1.6%
Procedure status	119,949	0.7%	112,736	1.3%
Drug groups primarily affecting the musculoskeletal system	227,337	1.3%	3	0.0%
Appliance	204,497	1.2%	49	0.0%
Drug groups and agents primarily acting on the skin	191,035	1.1%	1	0.0%
Foods, vitamins, electrolytes and inorganic salts	173,962	1.0%	12	0.0%
Substances, materials and objects	81,111	0.5%	57,773	0.7%
Family history	90,515	0.5%	39,265	0.5%
Drug groups primarily affecting the autonomic nervous system	105,199	0.6%	0	0.0%
Clinical examination	43,344	0.3%	42,012	0.5%
Prevention	67,804	0.4%	14,659	0.2%
[D]Symptoms, signs and ill-defined conditions	33,569	0.2%	45,772	0.5%
No relevant family history	64,863	0.4%	13,706	0.2%
Administrative values	48,841	0.3%	28,410	0.3%
Extinct cross-type investigation	62,901	0.4%	6,219	0.1%
Product bases and inactive substances	62,145	0.4%	0	0.0%

Level 2 - Categories	PCT1	%	PCT4	%
C/O - specified symptom findings	27,579	0.2%	28,027	0.3%
Vaccine, immunoglobulins, and antisera	40,735	0.2%	1,359	0.0%
Past medical history	26,898	0.2%	22,930	0.3%
Haematological agents	46,244	0.3%	0	0.0%
[V]Health status and contact with health services factors	27,734	0.2%	11,627	0.1%
Descriptors	32,610	0.2%	1,247	0.0%
Antineoplastic, immunosuppressant and immunostimulant	27,522	0.2%	0	0.0%
Prevention/screening administration	14,034	0.1%	7,707	0.1%
Assessment scales	16,328	0.1%	4,519	0.1%
Miscellaneous eye preparations	18,820	0.1%	1	0.0%
History/symptoms	10,957	0.1%	8,673	0.1%
Stoma appliance	18,085	0.1%	18	0.0%
Anaesthetics and medical gases	14,946	0.1%	0	0.0%
Extinct cross-type procedure	8,065	0.0%	4,319	0.1%
Accident	4,593	0.0%	6,846	0.1%
No significant medical history	5,236	0.0%	6,062	0.1%
Miscellaneous topical preparations	10,854	0.1%	0	0.0%
Monitoring administration	2,712	0.0%	7,074	0.1%
Tumour morphology	5,198	0.0%	4,085	0.0%
Drug groups primarily used in obstets, gynae+UT disorders	8,447	0.0%	0	0.0%
Reference documentation	3,881	0.0%	2,881	0.0%
Incontinence appliance	6,240	0.0%	0	0.0%
Human body structure	3,970	0.0%	2,528	0.0%
Education/welfare/health professions	4,130	0.0%	1,930	0.0%
Therapeutic prescription	1,437	0.0%	1,572	0.0%
Adverse reaction to drugs/medicines/biological substances	1,138	0.0%	3,120	0.0%
Indirect care procedures	2,867	0.0%	1,247	0.0%
Activities, functions and processes	1,318	0.0%	2,452	0.0%
Unknown	426,033	2.5%	37,052	0.4%
All others	36,934	0.2%	19,368	0.2%
Total	**16,957,920**		**8,459,723**	

Event_Date

This field represents the date the specified actual event occurred or the date the lab test was performed, broken down by year (where a year is considered to be from 1st of April through 31st of March of the **following** year).

Category	PCT1	%	PCT4	%
2002	4,958,302	29.2%	3,101,624	36.7%
2003	6,099,446	36.0%	2,683,694	31.7%
2004	5,900,172	34.8%	2,674,405	31.6%
Total	**16,957,920**		**8,459,723**	

Sex

This field represents the assigned gender to the given patient, where I indicates an invalid entry, and U indicates an unknown or indeterminate gender.

Category	PCT1	%	PCT4	%
F	10,286,907	60.7%	5,119,198	60.5%
I	0	0.0%	17	0.0%
M	6,670,847	39.3%	3,340,436	39.5%
U	166	0.0%	72	0.0%
Total	**16,957,920**		**8,459,723**	

Value 1

For particular Read codes, values are specified to further define the item being qualified and quantified. For example, if a blood pressure was recorded, the Read code would typically indicate whether the value was systolic or diastolic pressure, and this field would contain the associated blood pressure reading. If the value was not applicable, the value was always 0. Because of the varying nature of valid and expected values, it is difficult to assess the correctness of recorded data.

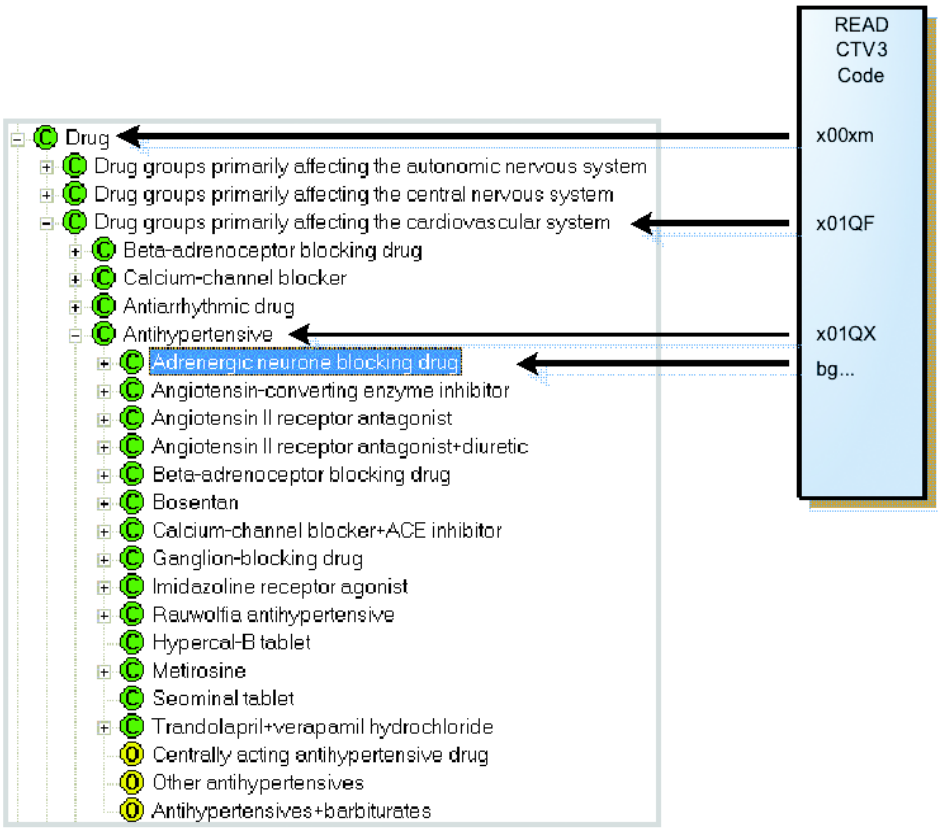
TRANSFORMATION – GP DATA

Read Code Versioning

Read code implementation was perhaps the most significant challenge, not only because of the various versions supported in different systems, but because of the changes encountered over the time period. (Read codes were converted into CTV3 terms and their associated CTV3 codes. Detailed instructions for mapping of Read codes between different versions are available from the NHS – Connecting for Health in the annual distribution of Read code tables.) Because of the non-numeric hierarchical nature of CTV3 Read codes, categorisations were made by grouping Read codes according to their location within the hierarchy. For example, suppose that the following records were encountered while evaluating GP data:

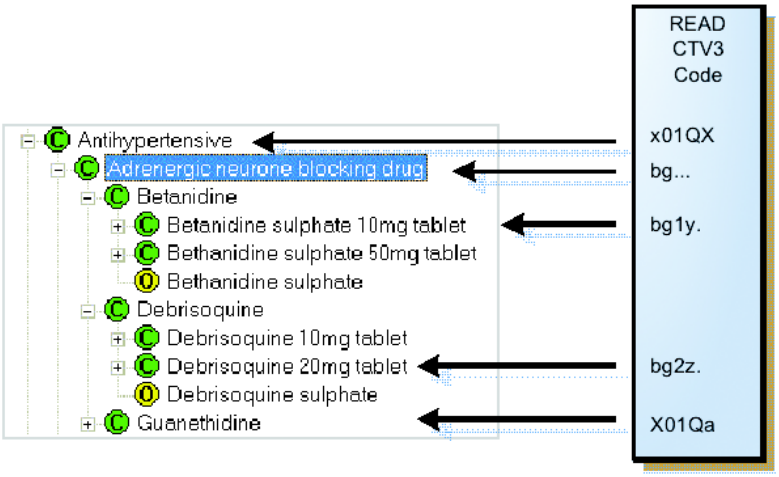
NHSNO	CTV3_CD	Description
0AB3EFA341	bg…	Adrenergic neurone blocking drug
0669A289192	bg1y.	Betanidine sulphate 10mg tablet
1012003910A	bg2z.	Debrisoquine 20mg tablet
338100092FA	X01Qa	Guanethidine

In the example below, to categorise the use of an antihypertensive drug as an indicator of the presence of a hypertensive condition (but not code to a specific antihypertensive drug), the code (X01QF) could be used to group all of these and all sub-codes below that group:



If a variable were defined to include detection of the use of **any** antihypertensive drug, a prescription of "Betanidine sulphate 10mg tablet" (coded as bg1y.) would trigger the indication of this variable as would the prescription of "Debrisoquine 20mg tablet" (coded as bg2z.). Although there exists a hierarchical relationship

between these read codes, there is no readily derivable lexical relationship. In order to provide this relationship in a usable format, a grouping table is supplied in the eMedia\Dictionary\GP folder so that each code can be referenced by its place in the hierarchy. For example, the Read code associated with antihypertensive drugs is "x01QX", which has no lexical relationship to "bg1y.". Looking in the eMedia\Dictionary\GP folder, in the ReadGrouped.csv file, this relationship is maintained as a straightforward lookup. Please see the eMedia\Dictionary\GP\ReadGrouped.doc for detailed instructions for use of this file.



Several inconsistencies were noted. Some practice data contained customised codes only valid internally to the practice implementing them – these were considered to be invalid, as were EMIS codes which could not be translated to their Read code equivalents. Records containing these codes were removed from consideration. All valid Read codes were mapped to their CTV3 equivalents for analysis.

When a blank was encountered, the value was padded with "." For example, if a simple code was encountered with a "bg", then this code was padded out to "bg…".

EMIS specific codes encountered in the Read code field were removed. To the extent possible, British National Formulary (BNF) codes were mapped to Read codes whenever BNF codes were encountered. Please see eMedia\Dictionary\GP\ReadFromBNF.csv for a detailed mapping table.

	PCT1	PCT4
All Records	16,989,555	22,897,392
Records in interval	16,981,931	8,523,293
Duplicates	1,223	5,897
Remaining	16,980,708	8,517,396
Removed Blanks	22,788	57,673
Remaining	16,957,920	8,459,723

SOCIAL SERVICES DATA

Social services data were excluded because of the inability to match records against existing data sources, either by NHS number, or unique identifiers such as birthdate and post code.

WAREHOUSE BUILDING

The "warehouse" is central to the concept of predictive modelling. The "warehouse" represents a common repository for consolidated information based on patients who exist in one or more available data sources. In essence, the warehouse is a compendium of the experiential data available.

This predictive model was developed using a warehouse approach that consolidated a central core of information regarding patients with an encounter within the initial two-year evaluation period. Only those PCTs that provided all four data sets were used in the creation of the warehouse. In order to define a true 4-dimensional model, each data set had to be submitted. Only data from PCT1 and PCT4 were utilised. Social Services data (as described earlier) were excluded after a review showed the number of patients with matching data across data sources could not be matched to data from Social Services data.

The warehouse consists of:

- IP file
- OP file
- A&E file
- GP file
- Member list

The first four files were assessed and transformed as described in the sections above.

MEMBER LIST

A member list ideally should be provided by each PCT and should include all patients registered with the PCT as of the date of data extraction, as well as relevant demographic (such as age and gender) and administrative (such as GP practice or active status) information. In the modelling performed, in the absence of member list provided, the member list was created including patients having an encounter within the first two years sampled, within any available data source.

Specific steps to creating the member list are listed below and are relatively straightforward.

INCLUSIONS

Develop a list of unique patients from each data set. All data sets are de-duplicated by NHS number, and the combined list de-duplicated. Only patients who had an encounter within the first two years evaluated were selected. The selection of others would be incompatible, since no previous history would be available. In the development of the Combined Model, PCT1 and PCT4 supplied data from all sources. The initial data were determined for **the interval of 2 years, beginning 1 April 2002 through 31 March 2004**. Initially, the following records were as shown:

	Records	
	PCT1	PCT4
IP	151,546	111,999
OP	1,223,546	323,568
A&E	195,582	70,406
GP	14,926,771[6]	5,829,098

Determine the matrix of membership appearing in each data set or combinations of data sets. This is useful because the predictive model developed is based on a similar matrix. The following table summarises the matrix of those patients existing in one or more data sets. (An asterisk indicates a record exists within this data set.)

6 This PCT submitted GP data in two separate files.

IP	OP	AE	GP	Patients	%
			*	245,341	43.6%
		*		28,254	5.0%
		*	*	33,447	5.9%
*				5,351	1.0%
*			*	11,302	2.0%
*		*		2,438	0.4%
*		*	*	7,427	1.3%
	*			32,771	5.8%
	*		*	74,118	13.2%
	*	*		7,183	1.3%
	*	*	*	25,365	4.5%
*	*			14,389	2.6%
*	*		*	38,635	6.9%
*	*	*		7,661	1.4%
*	*	*	*	29,029	5.2%

Determine the gender of that patient. Non-missing gender record associated with the maximum age from any one of the data sources was used.

Determine the age of that patient. If date of birth is provided, age can be calculated using the following formula:

Age = integer value ((end date of the 24-month period – date of birth)/365.25)

Otherwise the maximum non-missing age record from any one of the data sources was used.

EXCLUSIONS

All exclusions potentially apply to patients having records in different data sources for the period 1 April 2002 through 31 March 2004.

Exclude those patients that have a missing NHS number. NHS number is used as a unique patient identifier across the different data sources. If a patient does not have an NHS number across any of the datasets, it is not possible to link him/her to the records in the other files and calculate a risk score.

Exclude those patients with invalid or missing gender specification. Gender participates in the mathematical equation for calculation of risk score. The following table contains the number of records removed.

	Records	
	PCT1	PCT4
IP	9	0
OP	675	12
A&E	19	0
GP	126	161
Both male and female record	302	341

If a patient has a single record with invalid gender code in one data source and does not appear in any other data source, that patient would be excluded from the final list because of missing sex. Similarly a patient with contradicting gender records in different data sources would be excluded. However, if multiple records from one or more datasets for a specific patient existed, some with valid and some with invalid gender codes, that patient would be included in the final member list because of the records with valid sex code. The records specifying invalid gender were still included, but the gender was adjusted to reflect the correct gender as found in other datasets.

Exclude those patients with missing or invalid age. Age participates in the mathematical equations for calculation of risk score for all subpopulations. In the table below, each file was unduplicated by NHS number and the maximum age and sex from any file was retained. The maximum age was then determined from a combination of the available datasets. If the maximum age was either negative or missing, the patient was excluded from further consideration.

Patients With Missing or Invalid Ages	
PCT1	PCT4
34,759	4,033

In subsequent analysis, comparison between patients with missing and non-missing age shows a similar gender distribution, but reveals different utilisation rates in the year of prediction. For PCT1 utilisation rates were lower in the group with missing ages compared to the group with non-missing ages. For PCT2, utilisation rates were higher in the group with missing ages compared to the group with non-missing ages.

	Non-missing age		Missing age	
Gender distribution				
PCT1				
Male	161,648	46.9%	16821	48.4%
Female	183,085	53.1%	17938	51.6%
	344,733		34759	
PCT2				
Male	102,686	47.1%	1,967	48.8%
Female	115,292	52.9%	2,066	51.2%
	217,978		4,033	
Patients with IP Emergency Admission in the year following prediction				
PCT1	14,737	4.3%	704	2.0%
	344,733		34,759	
PCT2	11,199	5.1%	289	7.2%
	217,978		4,033	
Patients with OP visit in the year following prediction				
PCT1	85,922	24.9%	5,298	15.2%
	344,733		34,759	
PCT2	57,781	26.5%	1,174	29.1%
	217,978		4,033	
Patients with AE visit in the year following prediction				
PCT1	47,101	13.7%	2,561	7.4%
	344,733		34,759	
PCT2	27,705	12.7%	556	13.8%
	217,978		4,033	

Exclude those patients who are deceased. This is determined by examining the discharge method in inpatient data, where a "4" defines a patient who died during the inpatient spell, and "5" defines a stillbirth. A&E data were also examined, and entries from patients with an attendance_disposal_code of "10" (died in department) were removed, based on data from the first two years of the interval.

	Deceased Patients	
	PCT1	PCT4
IP	2,709	2,322
A&E	327	135

The final list of patients, by PCT, is as follows:

Unique Patients	
PCT1	PCT4
344,733	217,978
Total	**562,711**[7]

7 The combined population was split in two random samples: one for model development (281,094 patients) and one for model validation (281,617 patients).

MODEL SCORING

This section describes different types of variables considered in the modelling process, variables included in the model after the final variable selection, variables coding instructions, and instructions for calculation of risk score using the regression beta coefficients.

DEFINING AN OUTCOME

The model is designed to predict an outcome – in this case, the outcome of the model is an emergency admission with all of the following attributes:

- An admission method with a value of ("21", "22", "23", "24", "25," or "28"). While the HES dictionary does not define "25" as a valid code, PCT4 confirmed the usage of this code to indicate an emergency admission. The value of "25" was therefore considered to indicate the occurrence of an emergency admission. (PCT1 did not have any occurrences of this value.)
- A patient classification of ordinary admission ("1").
- An admission date within the 12 month period defined as the year of prediction.

TYPES OF VARIABLES CONSIDERED FOR PREDICTION

Different types of variables were determined in association with the source from which the values were extracted. The definition of the variable, where available, is delineated by IP (hospital inpatient data), OP (hospital outpatient data), A&E (accident and emergency data), and GP (general practitioner data extracted from physician data systems).

VARIABLES FROM IP DATA

Previous implementations of the PARR model included definition of variables based on IP hospital data, and are currently used in the PARR implementation and its Microsoft Access tool implementation. These variables are as follows:

- **Disease conditions**, based on ICD-10 codes grouped on a clinical basis. While some trusts placed diagnosis in chronological assigned order, others used this as an indicator of precedence. In the evaluation of diagnosis, equal weight was placed on the diagnosis, and subsequent diagnosis within the same spell, regardless of temporal occurrence within the spell.

- **Utilisation** (Emergency and Non-Emergency Admissions), based on ordinary admissions (Patient Classification=1), calculated one per patient per day. If multiple spells were specified within the same day, these were considered to be a single spell, with a total length of stay of 1 day, even if discharge occurred on the same day as admission.

- **Diagnostic cost groups/ hierarchical condition categories**, (implemented as predictor variables within the PARR model), were originally developed by researchers including Randall Ellis and Arlene Ash at Boston University, Brandeis University and Harvard Medical School to risk-adjust payments to managed care plans for the Medicare program in the United States[8]. The diagnosis for each patient is classified into a hierarchical diagnosis group (DxGroup) based on the seriousness of the patient's condition. DxGroups employs the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM), whereas, the United Kingdom uses the ICD-10. In order to utilise these ICD-9 based definitions, the ICD-10 codes were mapped to their ICD-9 version and to DxGroups. Patients with multiple diagnoses could be assigned to multiple DxGroups, by selecting the highest DXG group within the spell. DxGroups were then aggregated into payment groups, or DCGs (more than 20 in number). Final DXG group selection was determined by associated DXG group associated with the highest DCG group specified for a particular patient spell.

8 http://www.cms.hhs.gov/healthplans/rates/

- **Number day cases**, which specify the number of day case admissions and regular day admissions within a given time span.

Certain PARR variables could not be tested for inclusion within the Combined Model because of the use of encrypted information. Postcode or ward of census (2001) was either not available or encrypted – additionally, ethnicity was not consistently available for all patients. These demographic characteristics were excluded from analysis, including deprivation indices and population distributions upon which they are inherently based.

VARIABLES FROM OP DATA

OP variables were constructed based on an analysis of frequency, data-mining, clinical relevance and logic as defined below:

- **Number OP visits** was defined as one per patient per specialist per day, but only for those patients known to have attended defined by field **attended_or_dnad** with a value in ("5," "6", "05", "06", "Attended on time or, if late, before the relevant health care professional was ready to see the patient," "Arrived late, after the relevant health care professional was ready to see the patient, but was seen").
- **Number of distinct specialty visits**, based on the different types of specialists visited.
- **Other variables**, based on attendance, outcome of attendance, priority type, procedure status, reason for referral, and source of referral.

VARIABLES FROM A&E DATA

Similar to the OP variables, A&E variables were constructed based on an analysis of frequency, data mining, clinical relevance, and logic.

- **Number of A&E visits**, defined as one per patient per day.
- **Other variables**, based on length of stay in A&E, arrival mode, attendance category, attendance disposal, attendance initiator, location of incident, patient group, treatment, investigation, diagnosis, sub categorization of side of body, and sub categorization of site on body.

VARIABLES FROM GP DATA

GP variable creation presented unique challenges, due to different coding versions and specification within GP information systems. In order to level the values of the fields, all Read codes were resolved into their hierarchical groupings, and evaluated at different points in the hierarchy of membership. Six of these groupings (Appliances & Equipment for Level 4 grouping; Causes of Injury and Poisoning for Level 3 grouping; Clinical Findings for level 4 grouping: Context-dependent Categories for Level 4 grouping; Drugs for Level 3 grouping; and Operations, Procedures and Interventions for level 4 grouping) were further evaluated. Development of variables at these grouping levels allowed avoidance of the big variation in the use of single Read codes from different physicians, retaining at the same time the clinical meaning of the variables. (Please see the prior discussion concerning the use and resolution of Read code variables.)

Additional groups of variables were constructed and tested:

- **Polypharmacy**, defined as the number of unique drugs at Level 3 grouping per month.
- **Quantitative measurements**, including HbA1c values, alcohol consumption, body mass index (BMI), cholesterol, glomerular filtration rate (GFR; based on creatinine, age, sex, and assumed white ethnicity), forced expiratory volume (FEV), smoking status, blood pressure, and over 20 additional laboratory values. Both longitudinal and cross-sectional measurements were taken in consideration where data were available.

CHRONIC PROXY FLAGS

In the absence of chronic registries, proxy definitions were used to determine whether a specific patient could be considered to have a defined chronic condition. These proxy definitions were based on information from the IP data (ICD-10 diagnosis codes) and GP data (Read codes Level 4, grouping disorders; and Level 3 and level 4, grouping drugs). Specific chronic proxy flags were created for asthma, chronic obstructive pulmonary disease (COPD), coronary heart disease (CHD), congestive heart failure (CHF), depression, diabetes, hypertension, and

cancer. The performance of the proxy flags was then analysed by comparing frequency and chronic prevalence derived from registries later provided by one of the PCTs.

TIME LAG DETERMINATIONS

Each variable was created for five mutually exclusive time periods to determine the relevance of recency in each variable context with the assumption for a stronger effect associated with more recent time period:

- last 30 days
- last 30 to 90 days
- last 90 to 180 days
- last 180 to 365 days
- last 365 to 730 days

A version for the entire 2-year period of interest was created with each variable represented as either a flag, or, where appropriate, a count[9]. In the presence of low data frequencies (especially for most of the laboratory values), these time lags were then collapsed in 6- and 12-month periods. Different time lag versions for the same variable were included as independent variables in the variable selection procedure. While more than one time lag for the same variable could independently be selected in the model, if one was selected, the other time lags for the same variable were tested for their effect on the model performance. Please note the different time lagged versions described in the Variable Coding Instructions section.

VARIABLE CODING INSTRUCTIONS

LIST OF VARIABLES INCLUDED IN THE MODEL

The table below includes the variables selected for the Combined Model and the following information:

- **Short name of the variable used for coding purposes**. The first letters in the name (with some exceptions) indicate the type of variable or the data source the variable is coming from:
  AE = from A&E file
  GP = from GP file
  IP = from IP file
  OP = from OP file
  Ltc = chronic proxy flags
- **Description of the Variable**
- **Time Lag** for which the variable applies. For instance, if the 2-years-of-history period started on 1 April 2002 and ended on 31 March 2004, a Time Lag = "last 0 to 30 days" mandates that only records within the time interval 1 March 2004 through 31 March 2004 to be considered.
- **Beta Coefficients** produced from logistic regression in the process of model building.

Variable	Description	Beta Coefficient
Intercept		**-3.822847424**
agegrp0004	Age 0-4	0.289313618
agegrp1539	Age 15-39	0.385646264
agegrp4059	Age 40-59	0.373238463
agegrp6064	Age 60-64	0.630720996
agegrp6569	Age 65-69	0.481813417
agegrp7074	Age 70-74	0.507764968
agegrp7579	Age 75-79	0.813038432
agegrp8084	Age 80-84	0.959893138

9 In the final version of the model, count variables were converted into multiple binary variables to decrease the effect of extreme observations.

Variable	Description	Beta Coefficient
agegrp8589	Age 85-89	0.896645136
agegrp9094	Age 90-94	1.289601194
agegrp95pl	Age 95+	1.416839346
dem_gender	Gender[10]	0.01177781
AE_Invst01_m03_flg	AE visit - Investigation X-ray - last 90 to 180 days	0.216051313
AE_ArrAmb_m02_flg	AE visit - Arrived by ambulance - last 30 to 90 days	0.187349103
AE_DispRef_m01_flg	AE visit - Disposal to Specialist - last 0 to 30 days	0.632032184
AE_DxMed_m02_flg	AE visit - Medical DX (non-injury) - last 30 to 90 days	0.223716412
AE_DxMed_m12_flg	AE visit - Medical DX (non-injury) - last 365 to 730 days	0.321316757
AE_NumVisit1_m06_flg	1 AE visit - last 180 to 365 days	0.042763882
AE_NumVisit2_m06_flg	2 AE visits - last 180 to 365 days	0.290049439
AE_NumVisit3pl_m06_flg	3+ AE visits - last 180 to 365 days	0.507442635
Ltc_copd	COPD (LTC)	0.171100735
GP_dis47_y12	Psychoactive substance misuse disorder	0.54193793
GP_dis48_y12	Psychotic disorder	0.528176075
creatin_3_y02	Glomerular Filtration Rate Group 3	0.264393977
ChrCnt1_flg	1 (from 8[11]) LTC	0.119184904
ChrCnt2pl_flg	2+ (from 8) LTC	0.212972337
DisCnt7pl_flg	7+ distinct disorders (GP data)	0.096414136
GP_POLY_0104_123	1-4 unique drugs in any month - last 0 to 90 days	0.137302707
GP_POLY_0509_123	5-9 unique drugs in any month - last 0 to 90 days	0.388366204
GP_POLY_10pl_123	10+ unique drugs in any month - last 0 to 90 days	0.490961533
GP_drug36_m01_flg	Bronchodilator preparations - last 0 to 30 days	0.230925277
GP_drug36_m02_flg	Bronchodilator preparations - last 30 to 90 days	0.397601369
GP_drug36_m03_flg	Bronchodilator preparations - last 90 to 180 days	0.339967925
GP_drug36_m06_flg	Bronchodilator preparations - last 180 to 365 days	-0.403051621
GP_drug36_m12_flg	Bronchodilator preparations - last 365 to 730 days	-0.176615641
IP_DxMental_y12_flg	In-patient admission with diagnosis Mental illness - last 0 to 730 days	0.282235541
DiagCnt2_flg	2 distinct in-patient primary diagnosis (any episode) - last 0 to 730 days	0.132210548
DiagCnt3_flg	3 distinct in-patient primary diagnosis (any episode) - last 0 to 730 days	0.129497741
DiagCnt4pl_flg	4+ distinct in-patient primary diagnosis (any episode) - last 0 to 730 days	0.27788729
IP_util_EHRG_m01_flg	Emergency admission for impactable condition (HRG code) - last 0 to 30 days	0.482474391

10 dem_gender=1 if gender is female, 0 otherwise
11 Asthma, Diabetes, COPD, CAD, CHF, Hypertension, Depression, Cancer

Variable	Description	Beta Coefficient
IP_util_EHRG_m02_flg	Emergency admission for impactable condition (HRG code) - last 30 to 90 days	0.265806985
IP_util_EHRG_m03_flg	Emergency admission for impactable condition (HRG code) - last 90 to 180 days	0.260367409
IP_util_EHRG_m06_flg	Emergency admission for impactable condition (HRG code) - last 180 to 365 days	0.336849464
IP_util_E1pl_m01_flg	1+ Emergency admission - last 0 to 30 days	0.948115234
IP_util_E1_m02_flg	1 Emergency admission - last 30 to 90 days	0.476647042
IP_util_E2pl_m02_flg	2+ Emergency admissions - last 30 to 90 days	1.11137369
IP_util_E1_m03_flg	1 Emergency admission - last 90 to 180 days	0.346261242
IP_util_E2pl_m03_flg	2+ Emergency admissions - last 90 to 180 days	0.567774763
IP_util_E1_m06_flg	1 Emergency admission - last 180 to 365 days	0.20977492
IP_util_E2_m06_flg	2 Emergency admissions - last 180 to 365 days	0.352014497
IP_util_E3pl_m06_flg	3+ Emergency admissions - last 180 to 365 days	0.350301843
IP_util_E1_m12_flg	1 Emergency admission - last 365 to 730 days	0.312027413
IP_util_E2_m12_flg	2 Emergency admissions - last 365 to 730 days	0.32371827
IP_util_E3pl_m12_flg	3+ Emergency admissions - last 365 to 730 days	0.483011573
IP_util_EpisperE3pl_flg	Average number of episodes per Emergency admissions >=3	0.30864326
IP_HospOE	Observed/Expected ratio for rate of rehospitalisation for hospital of last admission	0.721855529
OP_NumVisit1_m01_flg	1 out-patient specialty visit - last 0 to 30 days	0.116311541
OP_NumVisit2_m01_flg	2 out-patient specialty visits - last 0 to 30 days	0.178716728
OP_NumVisit3pl_m01_flg	3+ out-patient specialty visits - last 0 to 30 days	0.291934635
OP_NumVisit1_m02_flg	1 out-patient specialty visit - last 30 to 90 days	0.150062538
OP_NumVisit2_m02_flg	2 out-patient specialty visits - last 30 to 90 days	0.151688397
OP_NumVisit3pl_m02_flg	3+ out-patient specialty visits - last 30 to 90 days	0.611030329
OP_NumVisit0105_m12_flg	1-5 out-patient specialty visits - last 365 to 730 days	0.182179996
OP_NumVisit0610_m12_flg	6-10 out-patient specialty visits - last 365 to 730 days	0.186734201
OP_NumVisit11pl_m12_flg	11+ out-patient specialty visits - last 365 to 730 days	0.364758425
OP_SrcRef5_m01_flg	OP visit - Source of referral not an Acc & Emergency - last 0 to 30 days	0.101656166
OP_SrcRef5_m02_flg	OP visit - Source of referral not an Acc & Emergency - last 30 to 90 days	0.322319293
Smoke_y02_Ltc_asth	Smoking status "yes" last 0-365 days multiplied by Asthma (LTC)	0.355326713
Ltc_copd_11pl_OP_visits_y12	11+ OP visits (last 0 to 730 days) multiplied by COPD (LTC)	-0.736470348

CODING INSTRUCTIONS

Legend	
NE	not equal
("a,b")	a or b
A - C	A or B or C
In	Included in the following list of values

Note: All variables are at patient level

IP VARIABLES

Variable	Time Lag	Type	Description	Coding
IP_DxMental_y12_flg	last 0 to 730 days	Flag	In-patient admission with diagnosis Mental illness	Primary_diagnosis or diag_2 - diag_6 = "F%". Exclude F10-F12, F13-F16,F70-F89,
DiagCnt2_flg	last 0 to 730 days	Flag	2 distinct in-patient primary diagnosis	2 unique 3-digit primary diagnosis (include any episode within the spell)
DiagCnt3_flg	last 0 to 730 days	Flag	3 distinct in-patient primary diagnosis	3 unique 3-digit primary diagnosis (include any episode within the spell)
DiagCnt4pl_flg	last 0 to 730 days	Flag	4+ distinct in-patient primary diagnosis	4+ unique 3-digit primary diagnosis (include any episode within the spell)
IP_util_EHRG_m01_flg	last 0 to 30 days	Flag	Emergency admission for impactable condition	1+ unique Admission_Date for Patient_Classification_Code="1" and Method_of_Admission_Code=("21","22","23","24", "25","28") and HRG3_code=(list, see eMedia\ Parameters\HRG3Impactable.csv)
IP_util_EHRG_m02_flg	last 30 to 90 days	Flag	Emergency admission for impactable condition	1+ unique Admission_Date for Patient_Classification_Code="1" and Method_of_Admission_Code=("21","22","23","24", "25","28") and HRG3_code=(list, see eMedia\ Parameters\HRG3Impactable.csv)
IP_util_EHRG_m03_flg	last 90 to 180 days	Flag	Emergency admission for impactable condition	1+ unique Admission_Date for Patient_Classification_Code="1" and Method_of_Admission_Code=("21","22","23","24", "25","28") and HRG3_code=(list, see eMedia\ Parameters\HRG3Impactable.csv)
IP_util_EHRG_m06_flg	last 180 to 365 days	Flag	Emergency admission for impactable condition	1+ unique Admission_Date for Patient_Classification_Code="1" and Method_of_Admission_Code=("21","22","23","24", "25","28") and HRG3_code=(list, see eMedia\ Parameters\HRG3Impactable.csv)
IP_util_E1pl_m01_flg	last 0 to 30 days	Flag	1+ Emergency admission	1+ unique Admission_Date for Patient_Classification_Code="1" and Method_of_Admission_Code=("21","22","23","24", "25","28")
IP_util_E1_m02_flg	last 30 to 90 days	Flag	1 Emergency admission	1 unique Admission_Date for Patient_Classification_Code="1" and Method_of_Admission_Code=("21","22","23","24", "25","28")
IP_util_E2pl_m02_flg	last 30 to 90 days	Flag	2+ Emergency admission	2+ unique Admission_Date for Patient_Classification_Code="1" and Method_of_Admission_Code=("21","22","23","24", "25","28")

Variable	Time Lag	Type	Description	Coding
IP_util_E1_m03_flg	last 90 to 180 days	Flag	1 Emergency admission	1 unique Admission_Date for Patient_Classification_Code="1" and Method_of_Admission_Code=("21","22","23","24","25","28")
IP_util_E2pl_m03_flg	last 90 to 180 days	Flag	2+ Emergency admission	2+ unique Admission_Date for Patient_Classification_Code="1" and Method_of_Admission_Code=("21","22","23","24","25","28")
IP_util_E1_m06_flg	last 180 to 365 days	Flag	1 Emergency admission	1 unique Admission_Date for Patient_Classification_Code="1" and Method_of_Admission_Code=("21","22","23","24","25","28")
IP_util_E2_m06_flg	last 180 to 365 days	Flag	2 Emergency admission	2 unique Admission_Date for Patient_Classification_Code="1" and Method_of_Admission_Code=("21","22","23","24","25","28")
IP_util_E3pl_m06_flg	last 180 to 365 days	Flag	3+ Emergency admission	3+ unique Admission_Date for Patient_Classification_Code="1" and Method_of_Admission_Code=("21","22","23","24","25","28")
IP_util_E1_m12_flg	last 365 to 730 days	Flag	1 Emergency admission	1 unique Admission_Date for Patient_Classification_Code="1" and Method_of_Admission_Code=("21","22","23","24","25","28")
IP_util_E2_m12_flg	last 365 to 730 days	Flag	2 Emergency admission	2 unique Admission_Date for Patient_Classification_Code="1" and Method_of_Admission_Code=("21","22","23","24","25","28")
IP_util_E3pl_m12_flg	last 365 to 730 days	Flag	3+ Emergency admission	3+ unique Admission_Date for Patient_Classification_Code="1" and Method_of_Admission_Code=("21","22","23","24","25","28")
IP_util_EpisperE3pl_flg	last 0 to 730 days	Flag	Average number of episodes per Emergency admissions >=3	(Total # of episodes for emergency admissions divided by Total # of emergency admissions) >=3
IP_HospOE	last 0 to 730 days	Continuous	Observed/Expected ratio for rate of rehospitalisation for hospital of last admission	Merge OE ratio list (see eMedia\Parameters\IP_HospOE.csv) to Provider_ID associated with the most recent Admission_date

OP VARIABLES

Variable	Time lag	Type	Description	Coding
OP_NumVisit1_m01_flg	last 0 to 30 days	Flag	1 out-patient specialty visit	1 unique (Date_of_attendance and Attended_or_DNAd = "5,6,05,06")
OP_NumVisit2_m01_flg	last 0 to 30 days	Flag	2 out-patient specialty visit	2 unique (Date_of_attendance and Attended_or_DNAd = "5,6,05,06")
OP_NumVisit3pl_m01_flg	last 0 to 30 days	Flag	3+ out-patient specialty visit	3+ unique (Date_of_attendance and Attended_or_DNAd = "5,6,05,06")
OP_NumVisit1_m02_flg	last 30 to 90 days	Flag	1 out-patient specialty visit	1 unique (Date_of_attendance and Attended_or_DNAd = "5,6,05,06")
OP_NumVisit2_m02_flg	last 30 to 90 days	Flag	2 out-patient specialty visit	2 unique (Date_of_attendance and Attended_or_DNAd = "5,6,05,06")
OP_NumVisit3pl_m02_flg	last 30 to 90 days	Flag	3+ out-patient specialty visit	3+ unique (Date_of_attendance and Attended_or_DNAd = "5,6,05,06")

Variable	Time Lag	Type	Description	Coding
OP_NumVisit0105_m12_flg	last 375 to 730 days	Flag	1-5 out-patient specialty visits	1-5 unique (Date_of_attendance and Attended_or_DNAd = "5,6,05,06")
OP_NumVisit0610_m12_flg	last 375 to 730 days	Flag	6-10 out-patient specialty visits	6-10 unique (Date_of_attendance and Attended_or_DNAd = "5,6,05,06")
OP_NumVisit11pl_m12_flg	last 375 to 730 days	Flag	11+ out-patient specialty visits	11+ unique (Date_of_attendance and Attended_or_DNAd = "5,6,05,06")
OP_SrcRef5_m01_flg	last 0 to 30 days	Flag	OP visit - Source of referral not an Acc & Emergency	Source_of_referral_code="5","05" and Attended_or_DNAd="5,6,05,06"
OP_SrcRef5_m02_flg	last 30 to 90 days	Flag	OP visit - Source of referral not an Acc & Emergency	Source_of_referral_code="5","05" and Attended_or_DNAd="5,6,05,06"

A&E VARIABLES

Variable	Time Lag	Type	Description	Coding
AE_Invst01_m03_flg	last 90 to 180 days	Flag	AE visit - Investigation X-ray	Att_Investigation_01_NatCode - Att_Investigation_06_NatCode="01"
AE_ArrAmb_m02_flg	last 30 to 90 days	Flag	AE visit - Arrived by ambulance	Att_ArrivalMode_NatCode="1,01"
AE_DispRef_m01_flg	last 0 to 30 days	Flag	AE visit - Disposal to Specialist	Att_Disposal_NatCode = ("3", "03")
AE_DxMed_m02_flg	last 30 to 90 days	Flag	AE visit - Medical DX (non-injury)	Att_Diag_01_NatCode=17-34 or Att_Diag_02_NatCode=17-34
AE_DxMed_m12_flg	last 365 to 730 days	Flag	AE visit - Medical DX (non-injury)	Att_Diag_01_NatCode=17-34 or Att_Diag_02_NatCode=17-34
AE_NumVisit1_m06_flg	last 180 to 365 days	Flag	1 AE visit	1 unique Att_Date_Arrival
AE_NumVisit2_m06_flg	last 180 to 365 days	Flag	2 AE visit	2 unique Att_Date_Arrival
AE_NumVisit3pl_m06_flg	last 180 to 365 days	Flag	3+ AE visit	3+ unique Att_Date_Arrival

GP VARIABLES

Variable	Time Lag	Type	Description	Read Code CTV3
GP_dis47_y12	last 0 to 730 days	Flag	Psychoactive substance misuse disorder	see eMedia\Parameters\GP_dis47.csv
GP_dis48_y12	last 0 to 730 days	Flag	Psychotic disorder	see eMedia\Parameters\GP_dis48.csv
DisCnt7pl_flg	last 0 to 730 days	Flag	7+ distinct disorders	Disorders are determined by selecting distinct level 4 Read codes within a Level 2 of 'X0003'. Please see the eMedia\Parameters\ReadGrouped.doc for a discussion regarding use of the eMedia\Dictionary\GP\ReadGrouped.csv file.
GP_drug36_m01_flg	last 0 to 30 days	Flag	Bronchodilator preparations	see eMedia\Parameters\GP_drug36.csv
GP_drug36_m02_flg	last 30 to 90 days	Flag	Bronchodilator preparations	see eMedia\Parameters\GP_drug36.csv
GP_drug36_m03_flg	last 90 to 180 days	Flag	Bronchodilator preparations	see eMedia\Parameters\GP_drug36.csv
GP_drug36_m06_flg	last 180 to 365 days	Flag	Bronchodilator preparations	see eMedia\Parameters\GP_drug36.csv
GP_drug36_m12_flg	last 365 to 730 days	Flag	Bronchodilator preparations	see eMedia\Parameters\GP_drug36.csv

QUANTITATIVE MEASUREMENTS

Creatin_3_y02 is based on GP records. Ethnicity is required for the calculation of GFR and since it was not provided, all calculations are based on measurements for ethnicity "white".

- Identify patients with Read code CTV3 for the time period specified.
- Determine the maximum measurement for the patient.
- Determine age as of the end of the time period specified.
- Use sex, age, and measurement recorded to assign patients to the group of interest.

Variable	Time Lag	Type	Description	Coding
Creatin_3_y02	last 0 to 365 days	Flag	Glomerular Filtration Rate group 3	X771Q XE2q5 44J3z XaERc XaETQ XaERX if sex=F and ((age<=25 and 110<value1<=200) or (25<age<=35 and 110<value1<=190) or (35<age<=45 and 100<value1<=180) or (45<age<=55 and 100<value1<=170) or (55<age<=65 and 90<value1<=170) or (65<age<=75 and 90<value1<=160) or (80<age and 90<value1<=160)) if sex=M and ((age<=25 and 150<value1<=260) or (25<age<=35 and 140<value1<=240) or (35<age<=45 and 130<value1<=230) or (45<age<=55 and 120<value1<=220) or (55<age<=65 and 120<value1<=220) or (65<age<=75 and 120<vlaue1<=210) or (80<age and 110<value1<=210))

POLYPHARMACY

Information from GP records is used to create the polypharmacy variables.

Variable	Time Lag	Type	Description	Coding
GP_POLY_0104_123	last 0 to 90 days	Flag	0-4 unique drugs (at Level 3 grouping) during any month in the last 0 to 90 days	See Instructions below
GP_POLY_0509_123	last 0 to 90 days	Flag	5-9 unique drugs (at Level 3 grouping) during any month in the last 0 to 90 days	See Instructions below
GP_POLY_10pl_123	last 0 to 90 days	Flag	10+ unique drugs (at Level 3 grouping) during any month in the last 0 to 90 days	See Instructions below

The variables are created as follows:

- Count number of unique drugs per month for a patient. Drugs are counted by selecting the number of distinct level 3 Read codes within a Level 1 of 'x00xm'. (Please see the eMedia\Parameters\ReadGrouped.doc for a discussion regarding use of the eMedia\Parameters\ReadGrouped.csv file.)

- If the maximum number of drugs taken for any month of the time period specified is 10 or more, a patient is assigned to group 10pl.

- Else If the maximum number of drugs taken for any month of the time period specified is 5-9, a patient is assigned to group 0509.

- Else If the maximum number of drugs taken for any month of the time period specified is 1-4, a patient is assigned to group 0104.

CHRONIC PROXY FLAGS (COPD, ASTHMA, DIABETES, CAD, CHF, HYPERTENSION, DEPRESSION, CANCER)

In the presence of information from chronic registries, patients with Long Term Conditions can be identified from the registries according to the Quality and Outcomes Framework (QOF) definitions for chronic conditions. In the absence of this information chronic proxy flags can be created as follows:

- Identify patients with ICD-10 diagnosis where the first 3 bytes of the ICD-10 code match the codes for every one of the conditions for any diagnosis field for the whole 2-year period. (eMedia\Parameters\IP_CHR_*.csv, where * represents ASTH, COPD, DIAB, CAD, CHF, HTEN, DEPR, and CNCR) from the IP file.

- Identify patients with Read code CTV3 from the list included in eMedia\Parameters\GP_CHR_*.csv for the conditions in the entire 2-year period from the GP file.

- Identify patients with a Read code CTV3 description contained at either a level3 or level 4 from the list included in eMedia\Parameters\DRUG_CHRL3_*.csv and eMedia\Parameters\DRUG_CHRL4_*.csv for each of the conditions in the entire 2-year period from the GP file.

- If a patient is flagged with any one of the three indicators, the patient is flagged with a LTC proxy flag appropriate to the condition tested.

Variable	Time Lag	Type	Description	Coding
ChrCnt1_flg	last 0 to 730 days	Flag	1 and only one Long Term Condition	ChrCnt1_flg is set to 1 if one and only one chronic condition (ASTH, CAD, CHF, CNCR, COPD, DEPR, HTEN), determined by having either a Read code as defined in eMedia\Parameters\GP_CHR_[ASTH,CAD,CHF,CNCR,COPD,DEPR, HTEN].csv (for GP derived data), or having a ICD-10 diagnosis in eMedia\Parameters\IP_CHR_[ASTH,CAD,CHF,CNCR,COPD,DEPR,HTEN].csv (for IP derived data)
ChrCnt2pl_flg	last 0 to 730 days	Flag	2 or more Long Term Conditions	ChrCnt2pl_flg is set to 1 if two or more distinct chronic conditions (ASTH, CAD, CHF, CNCR, COPD, DEPR, HTEN), determined by having either a Read code in eMedia\Parameters\GP_CHR_[ASTH,CAD,CHF,CNCR,COPD,DEPR, HTEN].csv (for GP derived data), or having a ICD-10 diagnosis in eMedia\Parameters\IP_CHR_[ASTH,CAD,CHF,CNCR,COPD,DEPR,HTEN].csv (for IP derived data)

INTERACTIONS

Variable	Description/Coding
Smoke_y02_Ltc_asth	Smoking status “yes” in last 0-365 days multiplied by Asthma(LTC). Smoking status is determined by one or more event_codes in ("137R.", "XE0oq", "Ub1tR", "XE0oi", "1372", Ub1tS", "1373", "Ub1tT", "1374", "Ub1tU", "1375", "Ub1tV", "Ub1tW", "1376", "137J.", "137H.", "XE0or", "XaBSp", "137Q.", "137P.", "XE0og", "XaIkY", "XaIkX", "XaIkW", "Ub1tI", "XaIuQ", "Ub1tJ", "Ub1tK", "137D.", "137Z.", "XE0oo", "Ub0p3", "Ub0p1", "Ub0p2", "137C.", "137G.", "137M.", "XaIIu", "XaItg") and not exists an event_code in ("XE0oo", "XaIQi", "XaIQm", "XaIQl", "XaIQk", "XaIQj", "XaIth", "XaIr7")
Ltc_copd_11pl_OP_visits_y12	(11+ unique Date_of_attendance and Attended_or_DNAd = "5,6,05,06") in last 0 to 730 days (OP) multiplied by COPD(LTC)

VARIABLE QUALITY CONTROL

It is recommended that after all variables are created, summary statistics be created for each of the variables to include mean, median, standard deviation (SD), minimum (Min), maximum (Max), quartiles (Q), and inter-quartile range (IQR). For binary variables (type flag, maximum value=1), the mean multiplied by the sample size results in a number of patients having the factor. The information could easily reveal variables with 0 frequencies, variables with extreme values, or any other anomalies caused either by incomplete or incorrectly entered data or coding mistakes. While it is relatively easy to identify variables with an unusual pattern, it is not always clear what caused the pattern and how to correct it.

For example, below are the summary statistics for predictor variables on the total validation sample and, separately, for PCT1 and PCT2.

TOTAL POPULATION

Variable	N	Mean	St_Dev	Min	Q1	Median	Q3	Max	IQR
AE_ArrAmb_m02_flg	281617	0.007	0.084	0	0	0	0	1	0
AE_DispRef_m01_flg	281617	0.009	0.095	0	0	0	0	1	0
AE_DxMed_m02_flg	281617	0.007	0.083	0	0	0	0	1	0
AE_DxMed_m12_flg	281617	0.035	0.185	0	0	0	0	1	0
AE_Invst01_m03_flg	281617	0.012	0.108	0	0	0	0	1	0
AE_NumVisit1_m06_flg	281617	0.067	0.250	0	0	0	0	1	0
AE_NumVisit2_m06_flg	281617	0.011	0.104	0	0	0	0	1	0
AE_NumVisit3pl_m06_flg	281617	0.004	0.064	0	0	0	0	1	0
agegrp0004	281617	0.072	0.258	0	0	0	0	1	0
agegrp1539	281617	0.355	0.479	0	0	0	1	1	1
agegrp4059	281617	0.257	0.437	0	0	0	1	1	1
agegrp6064	281617	0.047	0.213	0	0	0	0	1	0
agegrp6569	281617	0.042	0.201	0	0	0	0	1	0
agegrp7074	281617	0.036	0.187	0	0	0	0	1	0
agegrp7579	281617	0.031	0.173	0	0	0	0	1	0
agegrp8084	281617	0.024	0.154	0	0	0	0	1	0
agegrp8589	281617	0.013	0.115	0	0	0	0	1	0
agegrp9094	281617	0.007	0.082	0	0	0	0	1	0
agegrp95pl	281617	0.003	0.054	0	0	0	0	1	0
Ltc_copd	281617	0.008	0.087	0	0	0	0	1	0
ChrCnt1_flg	281617	0.155	0.362	0	0	0	0	1	0
ChrCnt2pl_flg	281617	0.053	0.223	0	0	0	0	1	0
creatin_3_y02	281617	0.009	0.097	0	0	0	0	1	0
dem_gender	281617	0.529	0.499	0	0	1	1	1	1
DiagCnt2_flg	281617	0.038	0.192	0	0	0	0	1	0
DiagCnt3_flg	281617	0.013	0.112	0	0	0	0	1	0
DiagCnt4pl_flg	281617	0.007	0.083	0	0	0	0	1	0

Variable	N	Mean	St_Dev	Min	Q1	Median	Q3	Max	IQR
DisCnt7pl_flg	281617	0.011	0.106	0	0	0	0	1	0
GP_dis47_y12	281617	0.003	0.056	0	0	0	0	1	0
GP_dis48_y12	281617	0.002	0.040	0	0	0	0	1	0
GP_drug36_m01_flg	281617	0.011	0.102	0	0	0	0	1	0
GP_drug36_m02_flg	281617	0.016	0.124	0	0	0	0	1	0
GP_drug36_m03_flg	281617	0.022	0.146	0	0	0	0	1	0
GP_drug36_m06_flg	281617	0.029	0.167	0	0	0	0	1	0
GP_drug36_m12_flg	281617	0.038	0.192	0	0	0	0	1	0
GP_POLY_0104_123	281617	0.183	0.387	0	0	0	0	1	0
GP_POLY_0509_123	281617	0.025	0.156	0	0	0	0	1	0
GP_POLY_10pl_123	281617	0.002	0.044	0	0	0	0	1	0
Smoke_y02_Ltc_asth	281617	0.008	0.092	0	0	0	0	1	0
Ltc_copd_11pl_OP_visits_y12	281617	0.001	0.027	0	0	0	0	1	0
IP_DxMental_y12_flg	281617	0.004	0.065	0	0	0	0	1	0
IP_HospOE	281617	0.191	0.386	0	0	0	0	1.648	0
IP_util_E1_m02_flg	281617	0.008	0.090	0	0	0	0	1	0
IP_util_E1_m03_flg	281617	0.012	0.110	0	0	0	0	1	0
IP_util_E1_m06_flg	281617	0.020	0.141	0	0	0	0	1	0
IP_util_E1_m12_flg	281617	0.034	0.181	0	0	0	0	1	0
IP_util_E1pl_m01_flg	281617	0.004	0.067	0	0	0	0	1	0
IP_util_E2_m06_flg	281617	0.003	0.053	0	0	0	0	1	0
IP_util_E2_m12_flg	281617	0.006	0.074	0	0	0	0	1	0
IP_util_E2pl_m02_flg	281617	0.001	0.029	0	0	0	0	1	0
IP_util_E2pl_m03_flg	281617	0.002	0.039	0	0	0	0	1	0
IP_util_E3pl_m06_flg	281617	0.001	0.028	0	0	0	0	1	0
IP_util_E3pl_m12_flg	281617	0.002	0.044	0	0	0	0	1	0
IP_util_EHRG_m01_flg	281617	0.001	0.026	0	0	0	0	1	0
IP_util_EHRG_m02_flg	281617	0.001	0.038	0	0	0	0	1	0
IP_util_EHRG_m03_flg	281617	0.002	0.047	0	0	0	0	1	0
IP_util_EHRG_m06_flg	281617	0.003	0.058	0	0	0	0	1	0
IP_util_EpisperE3pl_flg	281617	0.004	0.062	0	0	0	0	1	0
OP_NumVisit0105_m12_flg	281617	0.235	0.424	0	0	0	0	1	0
OP_NumVisit0610_m12_flg	281617	0.021	0.145	0	0	0	0	1	0
OP_NumVisit1_m01_flg	281617	0.050	0.219	0	0	0	0	1	0
OP_NumVisit1_m02_flg	281617	0.072	0.258	0	0	0	0	1	0
OP_NumVisit11pl_m12_flg	281617	0.005	0.071	0	0	0	0	1	0
OP_NumVisit2_m01_flg	281617	0.007	0.082	0	0	0	0	1	0
OP_NumVisit2_m02_flg	281617	0.015	0.121	0	0	0	0	1	0
OP_NumVisit3pl_m01_flg	281617	0.002	0.046	0	0	0	0	1	0
OP_NumVisit3pl_m02_flg	281617	0.002	0.048	0	0	0	0	1	0
OP_SrcRef5_m01_flg	281617	0.005	0.071	0	0	0	0	1	0
OP_SrcRef5_m02_flg	281617	0.008	0.088	0	0	0	0	1	0

PCT1

Variable	N	Mean	St_Dev	Min	Q1	Median	Q3	Max	IQR
AE_ArrAmb_m02_flg	172254	0.008	0.089	0	0	0	0	1	0
AE_DispRef_m01_flg	172254	0.010	0.102	0	0	0	0	1	0
AE_DxMed_m02_flg	172254	0.008	0.090	0	0	0	0	1	0
AE_DxMed_m12_flg	172254	0.042	0.201	0	0	0	0	1	0
AE_Invst01_m03_flg	172254	0.012	0.109	0	0	0	0	1	0
AE_NumVisit1_m06_flg	172254	0.074	0.262	0	0	0	0	1	0
AE_NumVisit2_m06_flg	172254	0.013	0.114	0	0	0	0	1	0
AE_NumVisit3pl_m06_flg	172254	0.006	0.074	0	0	0	0	1	0
agegrp0004	172254	0.087	0.282	0	0	0	0	1	0
agegrp1539	172254	0.374	0.484	0	0	0	1	1	1
agegrp4059	172254	0.246	0.431	0	0	0	0	1	0
agegrp6064	172254	0.041	0.199	0	0	0	0	1	0
agegrp6569	172254	0.037	0.189	0	0	0	0	1	0

Variable	N	Mean	St_Dev	Min	Q1	Median	Q3	Max	IQR
agegrp7074	172254	0.032	0.177	0	0	0	0	1	0
agegrp7579	172254	0.026	0.160	0	0	0	0	1	0
agegrp8084	172254	0.020	0.140	0	0	0	0	1	0
agegrp8589	172254	0.011	0.102	0	0	0	0	1	0
agegrp9094	172254	0.005	0.072	0	0	0	0	1	0
agegrp95pl	172254	0.003	0.052	0	0	0	0	1	0
Ltc_copd	172254	0.009	0.094	0	0	0	0	1	0
ChrCnt1_flg	172254	0.191	0.393	0	0	0	0	1	0
ChrCnt2pl_flg	172254	0.074	0.261	0	0	0	0	1	0
creatin_3_y02	172254	0.010	0.098	0	0	0	0	1	0
dem_gender	172254	0.529	0.499	0	0	1	1	1	1
DiagCnt2_flg	172254	0.034	0.182	0	0	0	0	1	0
DiagCnt3_flg	172254	0.011	0.104	0	0	0	0	1	0
DiagCnt4pl_flg	172254	0.006	0.075	0	0	0	0	1	0
DisCnt7pl_flg	172254	0.008	0.091	0	0	0	0	1	0
GP_dis47_y12	172254	0.003	0.052	0	0	0	0	1	0
GP_dis48_y12	172254	0.002	0.043	0	0	0	0	1	0
GP_drug36_m01_flg	172254	0.017	0.130	0	0	0	0	1	0
GP_drug36_m02_flg	172254	0.026	0.158	0	0	0	0	1	0
GP_drug36_m03_flg	172254	0.036	0.186	0	0	0	0	1	0
GP_drug36_m06_flg	172254	0.047	0.211	0	0	0	0	1	0
GP_drug36_m12_flg	172254	0.062	0.242	0	0	0	0	1	0
GP_POLY_0104_123	172254	0.299	0.458	0	0	0	1	1	1
GP_POLY_0509_123	172254	0.041	0.198	0	0	0	0	1	0
GP_POLY_10pl_123	172254	0.003	0.056	0	0	0	0	1	0
Smoke_y02_Ltc_asth	172254	0.011	0.103	0	0	0	0	1	0
Ltc_copd_11pl_OP_visits_y12	172254	0.001	0.031	0	0	0	0	1	0
IP_DxMental_y12_flg	172254	0.002	0.050	0	0	0	0	1	0
IP_HospOE	172254	0.166	0.363	0	0	0	0	1.36	0
IP_util_E1_m02_flg	172254	0.008	0.086	0	0	0	0	1	0
IP_util_E1_m03_flg	172254	0.011	0.105	0	0	0	0	1	0
IP_util_E1_m06_flg	172254	0.018	0.134	0	0	0	0	1	0
IP_util_E1_m12_flg	172254	0.030	0.169	0	0	0	0	1	0
IP_util_E1pl_m01_flg	172254	0.004	0.063	0	0	0	0	1	0
IP_util_E2_m06_flg	172254	0.002	0.047	0	0	0	0	1	0
IP_util_E2_m12_flg	172254	0.004	0.066	0	0	0	0	1	0
IP_util_E2pl_m02_flg	172254	0.001	0.025	0	0	0	0	1	0
IP_util_E2pl_m03_flg	172254	0.001	0.036	0	0	0	0	1	0
IP_util_E3pl_m06_flg	172254	0.001	0.024	0	0	0	0	1	0
IP_util_E3pl_m12_flg	172254	0.001	0.038	0	0	0	0	1	0
IP_util_EHRG_m01_flg	172254	0.001	0.025	0	0	0	0	1	0
IP_util_EHRG_m02_flg	172254	0.001	0.036	0	0	0	0	1	0
IP_util_EHRG_m03_flg	172254	0.002	0.046	0	0	0	0	1	0
IP_util_EHRG_m06_flg	172254	0.003	0.056	0	0	0	0	1	0
IP_util_EpisperE3pl_flg	172254	0.003	0.051	0	0	0	0	1	0
OP_NumVisit0105_m12_flg	172254	0.223	0.416	0	0	0	0	1	0
OP_NumVisit0610_m12_flg	172254	0.024	0.153	0	0	0	0	1	0
OP_NumVisit1_m01_flg	172254	0.050	0.219	0	0	0	0	1	0
OP_NumVisit1_m02_flg	172254	0.070	0.255	0	0	0	0	1	0
OP_NumVisit11pl_m12_flg	172254	0.007	0.081	0	0	0	0	1	0
OP_NumVisit2_m01_flg	172254	0.008	0.086	0	0	0	0	1	0
OP_NumVisit2_m02_flg	172254	0.016	0.126	0	0	0	0	1	0
OP_NumVisit3pl_m01_flg	172254	0.003	0.052	0	0	0	0	1	0
OP_NumVisit3pl_m02_flg	172254	0.003	0.054	0	0	0	0	1	0
OP_SrcRef5_m01_flg	172254	0.005	0.072	0	0	0	0	1	0
OP_SrcRef5_m02_flg	172254	0.008	0.087	0	0	0	0	1	0

PCT2

Variable	N	Mean	St_Dev	Min	Q1	Median	Q3	Max	IQR
AE_ArrAmb_m02_flg	109363	0.006	0.074	0	0	0	0	1	0
AE_DispRef_m01_flg	109363	0.007	0.082	0	0	0	0	1	0
AE_DxMed_m02_flg	109363	0.005	0.069	0	0	0	0	1	0
AE_DxMed_m12_flg	109363	0.025	0.155	0	0	0	0	1	0
AE_Invst01_m03_flg	109363	0.012	0.107	0	0	0	0	1	0
AE_NumVisit1_m06_flg	109363	0.055	0.228	0	0	0	0	1	0
AE_NumVisit2_m06_flg	109363	0.007	0.084	0	0	0	0	1	0
AE_NumVisit3pl_m06_flg	109363	0.002	0.043	0	0	0	0	1	0
agegrp0004	109363	0.048	0.214	0	0	0	0	1	0
agegrp1539	109363	0.325	0.469	0	0	0	1	1	1
agegrp4059	109363	0.275	0.446	0	0	0	1	1	1
agegrp6064	109363	0.057	0.232	0	0	0	0	1	0
agegrp6569	109363	0.050	0.217	0	0	0	0	1	0
agegrp7074	109363	0.042	0.202	0	0	0	0	1	0
agegrp7579	109363	0.038	0.190	0	0	0	0	1	0
agegrp8084	109363	0.032	0.175	0	0	0	0	1	0
agegrp8589	109363	0.018	0.131	0	0	0	0	1	0
agegrp9094	109363	0.009	0.096	0	0	0	0	1	0
agegrp95pl	109363	0.003	0.056	0	0	0	0	1	0
Ltc_copd	109363	0.006	0.077	0	0	0	0	1	0
ChrCnt1_flg	109363	0.099	0.298	0	0	0	0	1	0
ChrCnt2pl_flg	109363	0.020	0.139	0	0	0	0	1	0
creatin_3_y02	109363	0.009	0.095	0	0	0	0	1	0
dem_gender	109363	0.529	0.499	0	0	1	1	1	1
DiagCnt2_flg	109363	0.045	0.207	0	0	0	0	1	0
DiagCnt3_flg	109363	0.015	0.123	0	0	0	0	1	0
DiagCnt4pl_flg	109363	0.009	0.093	0	0	0	0	1	0
DisCnt7pl_flg	109363	0.016	0.126	0	0	0	0	1	0
GP_dis47_y12	109363	0.004	0.063	0	0	0	0	1	0
GP_dis48_y12	109363	0.001	0.034	0	0	0	0	1	0
GP_drug36_m01_flg	109363	0.000	0.000	0	0	0	0	0	0
GP_drug36_m02_flg	109363	0.000	0.000	0	0	0	0	0	0
GP_drug36_m03_flg	109363	0.000	0.000	0	0	0	0	0	0
GP_drug36_m06_flg	109363	0.000	0.000	0	0	0	0	0	0
GP_drug36_m12_flg	109363	0.000	0.000	0	0	0	0	0	0
GP_POLY_0104_123	109363	0.000	0.019	0	0	0	0	1	0
GP_POLY_0509_123	109363	0.000	0.000	0	0	0	0	0	0
GP_POLY_10pl_123	109363	0.000	0.000	0	0	0	0	0	0
Smoke_y02_Ltc_asth	109363	0.005	0.070	0	0	0	0	1	0
Ltc_copd_11pl_OP_visits_y12	109363	0.000	0.018	0	0	0	0	1	0
IP_DxMental_y12_flg	109363	0.007	0.084	0	0	0	0	1	0
IP_HospOE	109363	0.230	0.418	0	0	0	0	1.648	0
IP_util_E1_m02_flg	109363	0.009	0.096	0	0	0	0	1	0
IP_util_E1_m03_flg	109363	0.014	0.116	0	0	0	0	1	0
IP_util_E1_m06_flg	109363	0.024	0.152	0	0	0	0	1	0
IP_util_E1_m12_flg	109363	0.041	0.198	0	0	0	0	1	0
IP_util_E1pl_m01_flg	109363	0.005	0.073	0	0	0	0	1	0
IP_util_E2_m06_flg	109363	0.004	0.060	0	0	0	0	1	0
IP_util_E2_m12_flg	109363	0.007	0.085	0	0	0	0	1	0
IP_util_E2pl_m02_flg	109363	0.001	0.034	0	0	0	0	1	0
IP_util_E2pl_m03_flg	109363	0.002	0.043	0	0	0	0	1	0
IP_util_E3pl_m06_flg	109363	0.001	0.033	0	0	0	0	1	0
IP_util_E3pl_m12_flg	109363	0.003	0.052	0	0	0	0	1	0
IP_util_EHRG_m01_flg	109363	0.001	0.027	0	0	0	0	1	0
IP_util_EHRG_m02_flg	109363	0.002	0.041	0	0	0	0	1	0

Variable	N	Mean	St_Dev	Min	Q1	Median	Q3	Max	IQR
IP_util_EHRG_m03_flg	109363	0.002	0.049	0	0	0	0	1	0
IP_util_EHRG_m06_flg	109363	0.004	0.061	0	0	0	0	1	0
IP_util_EpisperE3pl_flg	109363	0.006	0.076	0	0	0	0	1	0
OP_NumVisit0105_m12_flg	109363	0.254	0.435	0	0	0	1	1	1
OP_NumVisit0610_m12_flg	109363	0.018	0.131	0	0	0	0	1	0
OP_NumVisit1_m01_flg	109363	0.050	0.219	0	0	0	0	1	0
OP_NumVisit1_m02_flg	109363	0.075	0.263	0	0	0	0	1	0
OP_NumVisit11pl_m12_flg	109363	0.003	0.052	0	0	0	0	1	0
OP_NumVisit2_m01_flg	109363	0.006	0.076	0	0	0	0	1	0
OP_NumVisit2_m02_flg	109363	0.013	0.113	0	0	0	0	1	0
OP_NumVisit3pl_m01_flg	109363	0.001	0.034	0	0	0	0	1	0
OP_NumVisit3pl_m02_flg	109363	0.001	0.037	0	0	0	0	1	0
OP_SrcRef5_m01_flg	109363	0.005	0.069	0	0	0	0	1	0
OP_SrcRef5_m02_flg	109363	0.008	0.089	0	0	0	0	1	0

INSTRUCTIONS FOR APPLYING BETA WEIGHTS

Risk score is derived at patient level using the predictor variables for that patient and respective beta coefficients included in Table 1 (Section "Variables Coding Instructions") to calculate log odds and then convert log odds to risk score. Natural logarithm is used in all calculations. Exp(-1 x log odds) is the exponential function, equivalent to e(-1log odds) .

Log odds = Intercept + Variable1 x Beta1 + Variable2 x Beta2 + Variable3 x Beta3 + … +
+ Variable68 x Beta68 + Variable69 x Beta69

Risk score = (1 / (1 + (exp(-1 x log odds)))) x 100

The risk score is on a scale of 0-100, with 100 indicating patients at highest risk of emergency admission in the 12 months following the 2 years of history.

Example:
Information for patient X and interim risk score calculations are presented below:

1. Multiply the value of the variable by the respective beta coefficient for that variable.

2. Sum the values for all variables to derive log odds.

Variable	Variable Value	Beta coefficient	Variable Value x Beta coefficient
Intercept		-3.822847424	-3.822847424
agegrp0004	0	0.289313618	0
agegrp1539	0	0.385646264	0
agegrp4059	1	0.373238463	0.373238463
agegrp6064	0	0.630720996	0
agegrp6569	0	0.481813417	0
agegrp7074	0	0.507764968	0
agegrp7579	0	0.813038432	0
agegrp8084	0	0.959893138	0
agegrp8589	0	0.896645136	0
agegrp9094	0	1.289601194	0
agegrp95pl	0	1.416839346	0
dem_gender	0	0.01177781	0
AE_ArrAmb_m02_flg	1	0.187349103	0.187349103
AE_DispRef_m01_flg	0	0.632032184	0

Variable	Variable Value	Beta coefficient	Variable Value x Beta coefficient
AE_DxMed_m02_flg	1	0.223716412	0.223716412
AE_DxMed_m12_flg	0	0.321316757	0
AE_Invst01_m03_flg	1	0.216051313	0.216051313
AE_NumVisit1_m06_flg	1	0.042763882	0.042763882
AE_NumVisit2_m06_flg	0	0.290049439	0
AE_NumVisit3pl_m06_flg	0	0.507442635	0
Ltc_copd	0	0.171100735	0
ChrCnt1_flg	0	0.119184904	0
ChrCnt2pl_flg	1	0.212972337	0.212972337
DisCnt7pl_flg	1	0.096414136	0.096414136
creatin_3_y02	0	0.264393977	0
GP_dis47_y12	0	0.54193793	0
GP_dis48_y12	0	0.528176075	0
GP_POLY_0104_123	0	0.137302707	0
GP_POLY_0509_123	1	0.388366204	0.388366204
GP_POLY_10pl_123	0	0.490961533	0
GP_drug36_m01_flg	0	0.230925277	0
GP_drug36_m02_flg	0	0.397601369	0
GP_drug36_m03_flg	0	0.339967925	0
GP_drug36_m06_flg	0	-0.403051621	0
GP_drug36_m12_flg	1	-0.176615641	-0.176615641
IP_DxMental_y12_flg	1	0.282235541	0.282235541
DiagCnt2_flg	0	0.132210548	0
DiagCnt3_flg	0	0.129497741	0
DiagCnt4pl_flg	1	0.27788729	0.27788729
IP_util_EHRG_m01_flg	0	0.482474391	0
IP_util_EHRG_m02_flg	0	0.265806985	0
IP_util_EHRG_m03_flg	0	0.260367409	0
IP_util_EHRG_m06_flg	0	0.336849464	0
IP_util_E1pl_m01_flg	1	0.948115234	0.948115234
IP_util_E1_m02_flg	0	0.476647042	0
IP_util_E2pl_m02_flg	0	1.11137369	0
IP_util_E1_m03_flg	1	0.346261242	0.346261242
IP_util_E2pl_m03_flg	0	0.567774763	0
IP_util_E1_m06_flg	1	0.20977492	0.20977492
IP_util_E2_m06_flg	0	0.352014497	0
IP_util_E3pl_m06_flg	0	0.350301843	0
IP_util_E1_m12_flg	0	0.312027413	0
IP_util_E2_m12_flg	0	0.32371827	0
IP_util_E3pl_m12_flg	0	0.483011573	0
IP_HospOE	0.899824278	0.721855529	0.649543131
IP_util_EpisperE3pl_flg	0	0.30864326	0
OP_NumVisit1_m01_flg	0	0.116311541	0
OP_NumVisit2_m01_flg	0	0.178716728	0
OP_NumVisit3pl_m01_flg	0	0.291934635	0
OP_NumVisit1_m02_flg	0	0.150062538	0
OP_NumVisit2_m02_flg	1	0.151688397	0.151688397
OP_NumVisit3pl_m02_flg	0	0.611030329	0
OP_NumVisit0105_m12_flg	0	0.182179996	0
OP_NumVisit0610_m12_flg	0	0.186734201	0
OP_NumVisit11pl_m12_flg	1	0.364758425	0.364758425
OP_SrcRef5_m01_flg	0	0.101656166	0
OP_SrcRef5_m02_flg	0	0.322319293	0
Smoke_y02_Ltc_asth	0	0.355326713	0
Ltc_copd_11pl_OP_visits_y12	0	-0.736470348	0
Sum			**0.971672967**

3. Exponentiate the sum:
   e(-1 x log odds) = exp(-1 x (0.971672967)) = 0.378449

4. Calculate a risk score on a scale from 0 to 1:
   Risk score = 1/ (1+ e(-1 x log odds)) = 1/(1+0.378449) = 0.725453

5. Transform the risk score on a scale from 0 to 100:
   Risk score = Risk score x 100 = 0.725453 x 100 = 72.54

RISK SCORE DISTRIBUTION

Risk score distribution for the total population, as well as certain segments of high risk patients included in the validation sample, is presented below:

Risk score	N	Mean	Min	Q1	Median	Q3	Max
Total Population	281,617	6	1	3	4	6	99
Top 250	250	81	71	75	80	86	99
Top 500	500	73	60	65	71	80	99
Top 1,000	1,000	63	49	53	60	71	99
Top 5,000	5,000	40	26	29	35	46	99
Top 10,000	10,000	30	18	21	26	35	99

Health Dialog UK Limited

Health Dialog is a leading provider of care management services, including disease management. It is one of the fastest growing privately held firms in the US, with a subsidiary company headquartered in Cambridge, UK. The firm's services include analytic services and telephonic care management support for individuals. Health Dialog helps individuals become more actively engaged in their healthcare and have more effective relationships with their clinicians.

King's Fund

The King's Fund is an independent charitable foundation working for better health, especially in London. We carry out research, policy analysis and development activities, working on our own, in partnerships, and through funding. We are a major resource to people working in health and social care, offering leadership development programmes; seminars and workshops; publications; information and library services; and conference and meeting facilities.

NYU – Robert F. Wagner Graduate School of Public Service

Established in 1938, the Robert F. Wagner Graduate School of Public Service offers advanced programs leading to the professional degrees of Master of Public Administration, Master of Urban Planning, Master of Science in Management, and Doctor of Philosophy. Through these rigorous programs, NYU Wagner educates the future leaders of public, non-profit, and health institutions as well as private organizations serving the public sector.