Additional power calculations

Due to the reduction in the study from three sites to one (see Supplementary Protocol), the available sample size was far smaller than originally intended. Using similar assumptions to the original protocol, except with various sample sizes, yielded the values of statistical power shown below. (Power calculations were undertaken in R v.3.5.0 using approach 2 of the power.nb.test function in the MKmisc library).

Fixed assumptions used in the power calculations included:

- average number of emergency inpatient admissions per patient per year = 0.6
- two-sided statistical test at a significance level of α = 5%
- an identifiable change in incidence rate ratio of admissions of 15%
- a study follow-up period of one year
- a negative binomial dispersion parameter θ = 1
- two control observations for each integrated care patient

Integrated care group sample size	Control group sample size	Resulting power
5,000	10,000	0.9998
2,500	5,000	0.9751
2,000	4,000	0.9391
1,500	3,000	0.8594
1,000	2,000	0.6985

Matching algorithm

Rather than include each potential control patient only once in the matching process, we constructed monthly ‘index dates’ at the midpoint of each month of the intervention year, 2014. We then used these as hypothetical intervention dates for the control patients. Thus, each control patient could provide up to 12 possible ‘observations’, with baseline variables and study endpoints recalculated with respect to each index date. This replication meant that a control patient could potentially contribute observations that were more closely matched to an integrated care patient than if they were only included once.

We subsequently excluded observations which did not meet the inclusion criteria used for the integrated care patients. Having assembled a dataset containing integrated care patients and up to 12 observations for each of the potential control patients, the dataset was stratified by risk score. At this point, it was found that 98.4% of the integrated care patients were within the top 7.5% of risk scores, whereas 35.7% of indexed control observations had risk scores this high. To improve computational efficiency during matching, the dataset was restricted to observations within the top 7.5% of risk scores, as this would retain nearly all the integrated care patients, whilst removing control observations which were unlikely to provide a good match.

From within this restricted set, individual matches were selected using genetic matching.29 Genetic matching is a computer-intensive search algorithm that can produce more closely matched control groups than traditional matching approaches. It uses machine learning – a ‘genetic’ algorithm41 – to search for parameter sets which maximise the p-values from paired t-tests and Kolmogorov-Smirnov tests, improving the balance of matched groups. We attempted to match two controls per intervention patient, with replacement.

We subsequently used a method of post-match adjustment – entropy balancing30 – to further improve the balance of the resulting matched dataset. This method makes use of Lagrange multipliers to identify weights for individual control observations that produce a near-perfectly balanced dataset, whilst maximising entropy42 – i.e. keeping the weights as near to their original values as possible.

Investigation of clustering

It was possible that clustering of patients within GP practices within health networks may have affected the results of the analyses. We were not able to use multilevel modelling approaches to control for this clustering. This is because the genetic matching and entropy balancing approaches used in this study to match integrated care patients to controls necessitate the use of weighted regression models. Currently, it is not possible to fit multilevel negative binomial models to weighted data in Stata (as of February 2019), and we are not aware of alternative software that can fit these models.

Instead, we ran OLS multilevel models of the same covariate specification as the main analyses, except that the health network was included not as a covariate (fixed effect) but instead as the highest level (level 3) random intercept, with GP practice and patient identifiers included as level 2 and level 1 random intercepts, respectively. The resulting residual intraclass correlations are shown below.

	Residual intraclass correlation coefficients		
	Health network (level 3)	GP practice (level 2)	Patient (level 1)
Emergency inpatient admissions	<0.0001	<0.0001	0.7246
Elective inpatient admissions	<0.0001	0.0031	0.6887
Inpatient bed days	<0.0001	<0.0001	0.8343
A&E attendances	<0.0001	0.0010	0.6728
Outpatient attendances	<0.0001	0.0007	0.5876
GP contacts	<0.0001	0.0617	0.6599

The coefficients shown above demonstrate that almost all the residual correlation was due to patient-level clustering. The only residual correlation of note at other levels was for GP contacts at the GP practice level (0.0617), but this was still very low. On this basis, using negative binomial models with cluster robust standard errors at the patient level seemed the best available approach for the main analyses.