Appendix 1. Populating the Data Mart

The Event Data Mart collected data from several different enterprise-wide information systems (the Operational DBs) in a large integrated delivery network (IDN) in New England. The IDN includes academic medical centers, community hospitals, specialty hospitals, and hospital-based and community-based ambulatory practices. The sources of data included access logs from a suite of clinical applications, user and clinic dictionaries from those systems, the human resources database, the patient registration data, the patient visit data, the credential management application database, the user authorization database, and the enterprise telecommunication directory. The data were collected as close as possible to the actual time of access and loaded into the Event Data Mart, a Microsoft SQL Server database where records from these different systems were linked.

In the Event Data Mart, each access to the EHR was stored in a table and termed as Record Access Event (RAE). An RAE, in this system, is the act of selecting a patient's record by a user from any application. The suite of clinical applications in use at this IDN check the enterprise master patient index (EMPI) before accessing the patient's record. From the EMPI, we are able to log real time registration information. The following data for each RAE also was logged at the time of access: user id, patient id, time of access, application being used, and network address of the computer from which the access was being made. In order to better understand the context around each RAE, we organized the other data into user, patient, and application categories, as outlined below:

User information: Users of clinical applications have entries in several operational databases. The data include where users work and live, with which sites and/or clinics they are affiliated, what their job title and job location are, which systems they have access to, and which associated roles are assigned to them within each of those systems.

Patient information: Patients have entries in registration and clinic scheduling systems. Included are data on where they live, where their scheduled visits are, their past encounters, their next-of kin contacts, their primary care physician, and their physical location and responsible service (for current inpatients) at the time a user accesses their EHR.

Application information: The application being used and where the computer is located at the time of access are used to understand the context of the RAE.

Appendix 2. Record Access Scenarios

We defined scenarios for appropriate access and suspicious access to EHRs with input from the privacy officers. The scenarios referred to as the baseline method, were used in selecting cases during the first round of training set construction.

The appropriate access scenarios involved a user accessing

(1) the record of a patient who had a recent visit to the user's clinic
(2) the record of a patient who was admitted to the user's care unit
(3) their own record (At this IDN, it is acceptable for a user to access their own record. At other institutions, accessing one's record is considered inappropriate.)
(4) the record of a patient for whom the user is a provider
(5) the record of a test patient

The suspicious access scenarios included the user accessing

(1) a co-worker's record
(2) the record of a patient who had not had a recent visit
(3) a VIP patient's record
(4) a family member's record
(5) the record of a patient who did not have a medical record number at the user's site
(6) a neighbor's record
(7) over 200 records in a day. In the latter scenarios, we filtered out events where there were reasons for appropriate access such as those involving a provider match, user's accessing their own records or those of test patients.

Appendix 3. Logistic Regression Results

Table A1. Univariate analysis of the training set. Odds ratio estimate and its 95% confidence interval are displayed. The total number of RAEs was 1,291 (643 suspicious RAEs and 648 appropriate RAEs).

Category	Feature	Odds Ratio	95% CI of Odds Ratio
User	Access after clinic hours	0.648	[0.447, 0.941]
	Access on clinic day	2.889	[2.073, 4.026]
	Over 200 accesses in a day	2.508	[1.431, 4.398]
	Hospital #1 employee	0.912	[0.712, 1.167]
	Hospital #2 employee	0.975	[0.778, 1.221]
	IDN employee	1.178	[0.897, 1.547]
	Nurse	0.698	[0.537, 0.906]
	Researcher	1.130	[0.771, 1.656]
Patient	Had recent visits	0.200	[0.138, 0.290]
	Is employee	1.322	[0.984, 1.775]
	VIP	0.883	[0.684, 1.139]
Relation	Is provider	0.147	[0.087, 0.249]
	Patient registered in user site	0.949	[0.736, 1.224]

Clinic visit match	0.287	[0.206, 0.399]
Care unit visit match	0.064	[0.040, 0.103]
Same city	1.577	[1.250, 1.989]
Works in same department	0.582	[0.392, 0.865]
Same family name	5.642	[4.374, 7.276]
Same street address	1.159	[0.804, 1.671]
Same zip code	0.469	[0.361, 0.610]