Supporting Information

Loukides et al. 10.1073/pnas.0911686107

SI Text

Related Work. In this section, we provide an extended discussion of related work that was not included in the main text owing to space limitations. The problem of preventing reidentification has been studied extensively by both the statistical disclosure control and database communities. The former community has proposed a number of methods that work by perturbing data. That is, they change the values of some attributes in such a way that they no longer correspond to real individuals (to prevent reidentification) but are statistically close to the original data (to preserve data utility). Popular methods, such as additive noise, data swapping, and synthetic data generation are popular perturbation-based methods (see refs. 1–3 for excellent surveys). These methods generate records that retain some aggregate statistics [e.g., the mean and correlations] and allow relatively accurate data mining models to be built [e.g., decision trees (4)]. However, these approaches are inappropriate for our scenario because the released records cannot be analyzed individually. Analysis of individual records is crucial for various clinical studies, such as determining the number of patient records that harbor a specific combination of ICD (International Classification of Diseases, 9th Revision, Clinical Modification) codes, which is important in epidemiology (5).

The database community has also proposed a number of works on the problem of preventing reidentification, which can be classified into two categories according to the data model they consider. The first category considers relational data (i.e., data in which a record has a fixed number of attributes that draw values from a specified domain). Many of these works are based on k-anonymity, a well-established principle proposed by Sweeney (6). A relational table satisfies this principle when each record is indistinguishable from at least $k - 1$ other records with respect to a set of potentially identifying attributes (termed quasi-identifiers or QIDs). K-anonymity is typically achieved by generalization, a process in which QID values are replaced by more general ones specified by a generalization model, or by suppression, a technique that removes values or records from anonymized data (6). Because both generalization and suppression distort data, constructing a table that satisfies k-anonymity should incur minimal distortion. The construction of such a table is modeled as an optimization problem, which can be solved using various search strategies (7–12). Our work is related to the aforementioned works because it uses generalization and suppression to prevent the association of an individual to their DNA sequence. However, we consider data that have very different semantics than that of relational data. More specifically, each record is associated with a set of ICD codes, and different records can harbor a variable number of ICD codes, which is typically much larger than the number of attributes in relational data. Thus, the data we consider cannot be dealt with adequately by the algorithms developed in (7–12), because they adopt crude generalization strategies (10) that limit the practical utility of data in biomedical applications (13), do not support for privacy requirements other than k-anonymity (e.g., they cannot be configured to deal with the single-visit case), and incur an excessive amount of information loss as well as large computational overhead when applied on high-dimensional data (11, 14, 15).

A number of other privacy principles that extend k-anonymity have also been proposed. Examples of such principles include l-diversity (17), (a, k)-anonymity (18), and tuple-diversity (14), which can be enforced by employing generalization and/or suppression (14, 16–18). All of these principles assume the existence of two types of attributes, QIDs and sensitive, and strengthen the protection provided by k-anonymity by additionally requiring values to follow a certain distribution with respect the sensitive attributes. These principles are not applicable to the scenario we consider, because we do not adopt this classification. ICD codes are partitioned into those that are potentially identifying or not, whereas the type of protection required for sensitive attributes is not suitable for protecting DNA sequences (e.g., DNA sequences are not susceptible to the "homogeneity" attack described in ref. 16).

Anonymizing transactional data (i.e., data in which a record, or *transaction*, is associated with a set of *items*) has been recently considered (14). Xu et al. (14) proposed (h, k, p)-coherence, a privacy principle that can prevent attackers with the knowledge of at most potentially identifying p items from linking an identified individual to fewer than k published transactions. In addition, (h, k, p)-coherence requires limiting the probability of associating an individual to a specified, sensitive item using a threshold h. An algorithm for enforcing (h, k, p)-coherence is also described by Xu et al. (15). This algorithm discovers all unprotected sets of items of minimal size and iteratively suppresses the item contained in the greatest number of those sets of items to satisfy (h, k, p)-coherence.

Although we consider transactional data (in our data an individual is associated with transaction, and ICD codes play the role of *items*), our work differs from that of Xu et al. (14) along four principal dimensions. First, our approach supports a much larger class of privacy requirements than that considered by Xu et al. (14). Although the approach described by Xu et al. (14) is effective at protecting all combinations of a certain number of ICD codes, potentially linkable combinations may involve certain ICD codes only and vary in size. In this case, the approach of Xu et al. (16) unnecessarily protects combinations of ICD codes, thereby excessively distorting data. Second, our approach allows the automatic extraction of privacy constraints from data, which is important because (i) it minimizes the effort of data owners that is required to determine which sets of ICD codes require protection. This can be extremely difficult in the setting we consider, because there is typically a very large number of ICD codes, and the number of combinations of ICD codes grows exponentially with this number. (ii) It provides the potential to improve data utility, because it avoids protecting combinations of items unnecessarily, which would harm utility because of the privacy/utility tradeoff (12). Third, the approach of Xu et al. (14) neglects specific data utility requirements, whereas our approach uses the notion of utility policy to guarantee that the released data will be of practical importance for the validation of genome-wide association studies (GWAS) when this utility policy is satisfied. Fourth, our algorithm differs from that proposed by Xu et al. (14) in that it uses suppression only when the privacy constraint cannot be satisfied by generalization alone. Therefore, our algorithm considers a significantly larger number of possible transformations to satisfy privacy constraints compared with that of Xu et al. (14), which offers greater opportunity for reducing the distortion incurred to anonymize data.

Anonymization Strategy. The following definition illustrates our anonymization strategy.

Definition 1 (anonymization strategy). *Let $\mathcal{I}$ be the set of all ICD codes that appear in the records of D. The anonymized dataset $\tilde{D}$ is derived from D by*

constructing a new set $\tilde{\mathcal{I}}$ such that: (i) each ICD code in $\mathcal{I}$ is uniquely mapped to an anonymized item $\tilde{i} \in \tilde{\mathcal{I}}$ that is a subset of $\mathcal{I}$, and (ii) $\mathcal{I} = (\cup_{m=1}^{|\tilde{\mathcal{I}}|} \tilde{i}_m) \cup \mathbf{S}$, where $|\tilde{\mathcal{I}}|$ denotes the size of $\tilde{\mathcal{I}}$ and $\mathbf{S}$ the set of suppressed ICD codes from $\mathcal{I}$ (i.e., those mapped to the empty subset of $\mathcal{I}$), and

replacing each ICD code i in D with the anonymized item this code has been mapped to if i has been mapped to a nonempty subset of $\mathcal{I}$, or suppressing i otherwise.

To illustrate the above definition, consider applying our anonymization strategy to the dataset shown in Fig. 1*A* (main text) to derive the anonymized dataset of Fig. 1*E* (main text). To create $\tilde{I}$, ICD codes 493.00, 493.01, and 493.02 are mapped to an anonymized item $\tilde{i}_1 = (493.00, 493.01, 493.02)$, the ICD codes 157.0, 157.1, 157.2, 157.3, and 157.9 are mapped to $\tilde{i}_2 = (157.0, 157.1, 157.2, 157.3, 157.9)$, and the ICD code 185 is mapped to $\tilde{i}_3 = (185)$. Because no ICD code is suppressed, we have $\tilde{I} = \tilde{i}_1 \cup \tilde{i}_2 \cup \tilde{i}_3$. Subsequently, the dataset shown in Fig. 1*E* (main text) is derived by replacing each of the ICD codes in D with the anonymized item $\tilde{i}_1, \tilde{i}_2$, or $\tilde{i}_3$ this ICD code has been mapped to.

Information Loss Measure. We model the amount of distortion caused by generalization and suppression of ICD codes using an information loss measure. We observe that the amount of distortion depends on (*i*) the number of profiles that contain the ICD code that is to be generalized (i.e., generalizing a frequent ICD code incurs a large amount of information loss), (*ii*) the number of ICD codes that are generalized together (i.e., mapping a large number of ICD codes to the same anonymized item incurs a large amount of information loss because it is difficult to distinguish the actual ICD codes using the anonymized item), and (*iii*) the semantic distance of the ICD codes that are generalized together (i.e., generalizing ICD codes together that are not closely related incurs a large amount of information loss). We propose the following measure that takes these factors into consideration.

Definition 2 (information loss measure). *Let $\mathcal{I}$ be the set of all ICD codes that appear in the records of D. The information loss for an anonymized item $\tilde{i}_m$ in $\tilde{I}$ is defined as*

$$IL(\tilde{i}_m) = \frac{2^{|\tilde{i}_m|} - 1}{2^M - 1} \times w(\tilde{i}_m) \times \frac{sup(\tilde{i}_m, \tilde{\mathcal{D}})}{N}$$

where $|\tilde{i}_m|$ denotes the number ICD codes that are mapped to $\tilde{i}_m$, M is the number of ICD codes in $\mathcal{I}$, N is the number of records in D, $\tilde{\mathcal{D}}$ is the anonymized version of D, $sup(\tilde{i}_m, \tilde{\mathcal{D}})$ represents the number of records of $\tilde{\mathcal{D}}$ that are subsets of $\tilde{i}_m$, and $w : \tilde{\mathcal{I}} \rightarrow [0, 1]$ is a function assigning a weight to $\tilde{i}_m$ based on the semantical distance of the ICD codes it contains. We compute this distance following Xu et al. (19) and using the ICD coding hierarchy.

To illustrate how the above measure can be computed, consider the ICD codes 493.00, 493.01, and 493.02 in the dataset of Fig. 1*A* (main text), which are mapped to the anonymized item $\tilde{i}_1 = (493.00, 493.01, 493.02)$ in the data of Fig. 1*E* (main text), and a weight of 0.375. Using *IL*, we can compute the information loss of this generalization as $IL(\tilde{i}_1) = \frac{2^3 - 1}{2^9 - 1} \times 0.375 \times \frac{6}{9} \approx 0.0034$. Similarly, using the same weight, we can compute the information loss incurred by mapping the ICD codes 157.0, 157.1, 157.2, 157.3, and 157.9 to the anonymized item $\tilde{i}_2 = (157.0, 157.1, 157.2, 157.3, 157.9)$ as $IL(\tilde{i}_2) = 0.0072$. Notice that the latter generalization incurs a higher amount of information loss, because $\tilde{i}_2$ comprises a larger number of ICD codes than that of $\tilde{i}_1$ and is associated with a larger number of patients in the anonymized dataset.

Pseudocode for PPE and UGACLIP. In what follows, we provide the pseudocode for the Privacy Policy Extraction (PPE) and the Utility-Guided Anonymization of CLInical Profiles (UGACLIP) algorithms.

Algorithm 1 illustrates the PPE algorithm. PPE takes as input a set of records derived from the original database D after applying a filtering condition $\mathcal{F}$ on D, as well as the anonymization threshold k, and returns a privacy policy $\mathcal{P}$ that can subsequently be provided as input to UGACLIP.

Algorithm 1: the PPE Algorithm.

1. **procedure** PPE ($\mathcal{D} \leftarrow$ apply $\mathcal{F}$ to D, threshold k)
2. Sort the records of $\mathcal{D}$ in descending order based on the number of ICD codes they contain
3. **for** (each record r_i of $\mathcal{D}$, $i = 1, \ldots, N$)
4. **for** (each record r_j of $\mathcal{D}$, $j = i + 1, \ldots, N$)
5. **if** (all ICD codes of r_j that are contained in r_i)
6. Discard record r_j from $\mathcal{D}$
7. **end if**
8. **end for**
9. **if** (at least k records of $\mathcal{D}$ contain the ICD codes of r_i)
10. Discard record r_i from D
11. **end if**
12. **end for**
13. Return a privacy policy $\mathcal{P}$ containing a privacy constraint for each remaining record in $\mathcal{D}$
14. **end procedure**

The applied filtering condition $\mathcal{F}$ can be modeled as a simple selection and projection operation on D, which combines ICD codes across different records. In this article, we consider two different filtering conditions: (*i*) $\mathbf{f_1}$ (the single-visit case), which, for each patient and each of their service dates, retrieves the set of ICD codes that the patient was diagnosed with; and (*ii*) $\mathbf{f_2}$ (the all-visits case), which retrieves the set of ICD codes that were assigned to a patient on all visits collectively. It should also be noted that PPE allows data owners to use different filtering conditions according to their expectations regarding which ICD codes are potentially identifying.

Given the sets of ICD codes (denoted here as *records*) that were extracted from database D by using $\mathcal{F}$, the PPE algorithm sorts them in decreasing order according to their size and stores them as a dataset $\mathcal{D}$. Then, PPE iterates over each individual record in $\mathcal{D}$ and deletes a record that is a superset of another record in $\mathcal{D}$. This results in retaining the sets of ICD codes which, when protected in D, will also lead to the protection of all their subsets (including those captured by the deleted records of $\mathcal{D}$). After removing the subsets of a given record from D, the PPE algorithm checks whether the corresponding combination of ICD codes is already protected in D (i.e., there are at least k records that harbor this combination of ICD codes in $\mathcal{D}$). In this case, this record is deleted. After iterating over all records of the dataset, PPE returns a privacy policy $\mathcal{P}$ that contains a privacy constraint for each of the remaining records in $\mathcal{D}$.

UGACLIP algorithm. Algorithm 2 demonstrates the operation of UGACLIP. Given the original database D, the anonymization threshold k, the privacy policy $\mathcal{P}$, and the utility policy $\mathcal{U}$, UGACLIP anonymizes D in a series of steps. First, the anonymized database $\tilde{D}$ is initialized to D. Then, UGACLIP selects the privacy constraint p that is currently associated with the most number of patients in $\tilde{D}$ and tries to satisfy it. To achieve this, UGACLIP first checks whether p can still be generalized according to the utility policy $\mathcal{U}$. If this is the case, it selects the least frequent ICD code i from p and finds the utility constraint u that contains it. Next, it checks whether u has at least two ICD codes (this ensures that generalizing i is possible) and finds the ICD code i' from u with which i can be generalized in a way that minimizes the *IL* measure of definition 2. Subsequently, $\tilde{D}$ is updated to reflect this generalization. In the case that u contains

only one ICD code and i appears in fewer than the minimum required number of records in $\tilde{D}$, UGACLIP suppresses i. The suppression of i deletes this ICD code from every record in $\tilde{D}$ that contains it. When the above steps are insufficient to satisfy the privacy constraint p, UGACLIP suppresses all items in p to satisfy it. UGACLIP repeats the same process to satisfy all privacy constraints in the privacy policy $\mathcal{P}$. When all privacy constraints are satisfied, the anonymized database $\tilde{D}$ is returned as a result.

Relative Error. To test the effectiveness of our method in generating anonymizations that support clinical studies focusing on case counts, we used the *relative error* (*RE*) score, a widely adopted data utility criterion that measures the difference in the accuracy of answering queries on the original and anonymized data. The queries we considered involved counting the number of patients diagnosed with a certain set of ICD codes. Such queries can be modeled as follows:

Q: SELECT COUNT(*)
FROM dataset
WHERE $ICD_1 \in$ dataset and $ICD_2 \in$ dataset and .... and $ICD_q \in$ dataset

Algorithm 2: the UGACLIP algorithm.

1. **procedure** UGACLIP (database D, threshold k, privacy policy $\mathcal{P}$, utility policy $\mathcal{U}$)
2. $\tilde{D} \leftarrow D$
3. **while** (privacy policy $\mathcal{P}$ is not satisfied)
4. $p \leftarrow$ privacy constraint currently associated with most patients
5. **while** (privacy constraint p is not satisfied $\wedge$ p can still be generalized given $\mathcal{U}$)
6. $i \leftarrow$ the least frequent ICD code in $\tilde{D}$ from p
7. $u \leftarrow$ the utility constraint from $\mathcal{U}$ that contains i
8. **if** (u contains at least two ICD codes)
9. Generalize i with another ICD code i' in u such that the *IL* of the resulting anonymized item is minimum
10. Update records of $\tilde{D}$ to reflect the new generalization
11. **else if** (i appears in fewer than k records of $\tilde{D}$)
12. Suppress i
13. Update records of $\tilde{D}$ to reflect the suppression of i
14. **end if**
15. **end while**
16. **if** (p is not satisfied)
17. Suppress p
18. Update the records of $\tilde{D}$ to reflect the suppression of p
19. **end if**
20. **end while**
21. Return the anonymized database $\tilde{D}$
22. **end procedure**

Assume that $a(Q)$ is the answer of applying Q to the original dataset D, which can be obtained by counting the number of records in D that contain a certain set of ICD codes. When the same query Q is applied on the anonymized data $\tilde{\mathcal{D}}$, we obtain an estimated answer $e(Q)$, because the anonymized items may not allow distinguishing the actual ICD codes a patient has. To compute $e(Q)$, we first need to compute the probability a patient is diagnosed with the requested ICD code(s). The latter probability can be computed as $\prod_{r=1}^{q} p(i_r)$, where $p(i_r)$ is the probability of mapping an ICD code i_r in query Q to an anonymized item $\tilde{i}_m$, assuming that $\tilde{i}_m$ can include any possible subset of the ICD codes mapped to it with equal probability (i.e., uniform distribution) and that there exist no correlations among the anonymized items. The estimated answer $e(Q)$ is then derived by summing the corresponding probabilities across the records of $\tilde{\mathcal{D}}$.

Once $a(Q)$ and $e(Q)$ are known, we compute *RE* for Q as $RE(Q) = |a(Q) - e(Q)|/a(Q)$. Intuitively, the lower the *RE* score for a workload of queries, the higher the quality of the anonymization method because it can more accurately compute from the anonymized clinical profiles the number of patients' records that correctly answer the query. Given a set of queries similar to Q, we use both the mean and the SD of the *RE* for these queries to capture how accurately these queries can be computed using the anonymized dataset.

Utility Policies. Table S1 illustrates the utility policies specified for different values of k between 2 and 25. The associations between diseases used in these utility policies and ICD codes are illustrated in Table S2. Each disease corresponds to a utility constraint, which is modeled as a set of ICD codes. For clarity, we only present ICD codes that are contained in the used datasets.

Tables S3–S5 illustrate the utility policies for the single-visit case and for $k = 2$, $k = 10$, and $k = 25$, respectively. The all-visits case is illustrated in Table S6–S9 for $k = 2$, $k = 5$, $k = 10$, and $k = 25$, respectively. Notice that some diseases that appear in the utility policy for $k = 2$ do not appear in the corresponding policies for larger values of k. This is because diseases that do not appear at least k times in the datasets are not included in a utility policy, because they must be suppressed to satisfy the utility policy.

Finally, we provide details regarding the amount of suppression performed by UGACLIP in both the single-visit and the all-visits case. Fig. S1*A* reports the number of suppressions (i.e., $\sum_{\forall \tilde{i}_m \in \mathbf{S}} sup(\tilde{i}_m, \tilde{\mathcal{D}})$, where $sup(\tilde{i}_m, \tilde{\mathcal{D}})$ is as defined in definition 2), whereas Fig. S1*B* illustrates the number of distinct ICD codes that are suppressed (i.e., the size of the set of suppressed items **S**). The results of Fig. S1 *A* and *B* correspond to both used datasets and verify that the amount of suppression performed by UGACLIP is small. This because suppression is used only when using generalization would violate the specified privacy policy, whereas the domain size (i.e., the space of items that can potentially be suppressed) is in the order of thousands. We also note that ACLIP performs no suppression, because it works by applying generalization in a way that minimizes information loss and without taking utility constraints into account.

Support of Clinical Case Counts for the All-Visits Case. The results for the all-visits case for the Vanderbilt Native Electrical Conduction dataset (*VNEC*) and Vanderbilt Native Electrical Conduction Known Controls dataset ($VNEC_{KC}$) datasets are reported in Figs. S2 *A* and *B* respectively.

1. Adam NR, Wortman JC (1989) Security conrol methods for statistical databases. *ACM Comput Surv* 21:515–556.
2. Aggarwal CC, Yu PS, eds (2008) A survey of randomization methods for privacy-preserving data mining. *Privacy-Preserving Data Mining: Models and Algorithms* (Springer, New York), pp 267–289.
3. Willenborg L, De Waal T (1996) *Statistical Disclosure Control in Practice. Lecture Notes in Statistics* (Springer, New York)Vol Vol 111.
4. Agrawal R, Srikant R (2000) Privacy-preserving data mining. *ACM SIGMOD Record* 29: 439–450.
5. Marsden-Haug N, et al. (2007) Code-based syndromic surveillance for influenzalike illness by International Classification of Diseases, Ninth Revision. *Emerg Infect Dis* 13: 207–216.
6. Sweeney L (2002) k-anonymity: A model for protecting privacy. *Int J Uncertainty Fuzziness Knowledge-based Systems* 10:557–570.

7. Sweeney L (2002) Achieving k-Anonymity privacy protection using generalization and suppression. *Int J Uncertainty Fuzziness Knowledge-Based Systems* 10:571–588.
8. LeFevre K, DeWitt DJ, Ramakrishnan R (2006) Mondrian multidimensional k-anonymity. *Proceedings of the International Conference on Data Engineering*, eds Liu L, Reuter R, Whang KY, Zhang J (IEEE Computer Society, Los Alamitos), p 25.
9. Ghinita G, Karras P, Kalnis P, Mamoulis N (2007) Fast data anonymization with low information loss. *Proceedings of the International Conference on Very Large Data Bases*, eds Koch C et al., (ACM, New York), pp 758–769.
10. LeFevre K, DeWitt DJ, Ramakrishnan R (2005) Incognito: Efficient full-domain k-anonymity. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, ed Ozcan F (ACM, New York), pp 49–60.
11. Loukides G, Shao J (2007) Clustering-based k-anonymisation algorithms. *Proceedings of the International Conference on Database and Expert Systems Applications*, eds Wagner R, Revell N, Pernul G (Springer, Heidelberg), pp 761–771.
12. Loukides G, Shao J (2007) Capturing data usefulness and privacy protection in k-anonymisation. *Proceedings of the ACM Symposium on Applied Computing*, eds Cho Y et al., (ACM, New York), pp 370–374.
13. Rogers JE (2006) Quality assurance of medical ontologies. *Methods Inf Med* 45:267–274.
14. Xu Y, Wang K, Fu AWC, Yu PS (2008) Anonymizing transaction databases for publication. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, eds Li Y, Liu B, Sarawagi S (ACM, New York), pp 767–775.
15. Aggarwal CC (2005) On k-anonymity and the curse of dimensionality. *Proceedings of the International Conference on Very Large Data Bases*, eds Bohm K et al., (ACM, New York), pp 901–909.
16. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M (2006) l-diversity: Privacy beyond k-anonymity. *Proceedings of the International Conference on Data Engineering*, eds Liu L, Reuter R, Whang KY, Zhang J (IEEE Computer Society, Los Alamitos), p 24.
17. Li N, Li T, Venkatasubramanian S (2007) t-Closeness: privacy beyond k-anonymity and l-diversity. *Proceedings of the International Conference on Data Engineering*, eds Chirkova R, Dogac A, Oszu T, Sellis T (IEEE Computer Society, Los Alamitos, USA), pp 106–115.
18. Wong RC, Li J, Fu AW, Wang K (2006) (a, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, eds Ungar L, Craven M, Gunopulos D, Eliassi-Rad T (ACM, New York, USA), pp 754–759.
19. Xu J, et al. (2006) Utility-based anonymization using local recoding. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, eds Ungar L, Craven M, Gunopulos D, Eliassi-Rad T (ACM, New York, USA), pp 785–790.

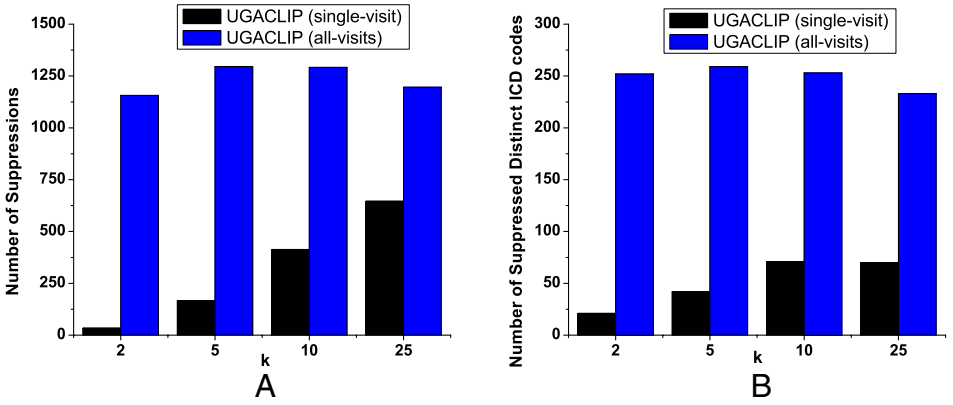

Fig. S1. Amount of suppression performed by UGACLIP for the single-visit and all-visits cases. (*A*) number of suppressions vs. *k*. (*B*) Number of distinct ICD codes that are suppressed vs. *k*.

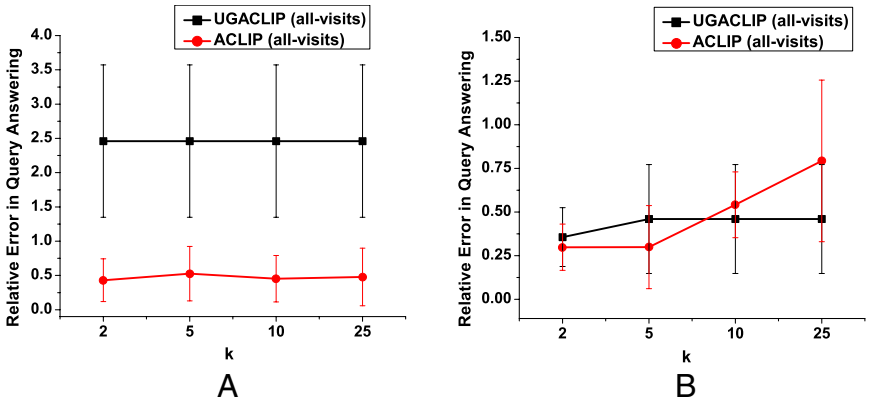

Fig. S2. *Relative error* in query answering for the all-visits case and for (*A*) *VNEC* and (*B*) $VNEC_{KC}$. Points correspond to the mean *RE*, and error bars are of 1 SD.

Table S1. Diseases contained in the utility policy for values of *k* between 2 and 25 (✓ denotes that a disease is contained in a utility policy)

Disease	$k = 2$	$k = 5$	$k = 10$	$k = 25$
Asthma	✓	✓	✓	✓
Attention deficit with hyperactivity	✓	✓	✓	
Bipolar I disorder	✓	✓	✓	✓
Bladder cancer	✓	✓	✓	✓
Breast cancer	✓	✓	✓	✓
Coronary disease	✓	✓	✓	✓
Dental caries	✓	✓	✓	✓
Diabetes mellitus type 1	✓	✓	✓	✓
Diabetes mellitus type 2	✓	✓	✓	✓
Lung cancer	✓	✓	✓	✓
Major depressive disorder	✓			
Pancreatic cancer	✓	✓	✓	✓
Platelet phenotypes	✓	✓	✓	✓
Preterm birth	✓	✓	✓	✓
Prostate cancer	✓	✓	✓	✓
Psoriasis	✓	✓	✓	✓
Renal cancer	✓	✓	✓	
Schizophrenia	✓	✓	✓	✓
Sickle-cell disease	✓	✓		✓

Table S2. Diseases used as utility constraints and their corresponding sets of ICD codes

Disease	Set of ICD codes
Asthma	{493.00, 493.01, 493.02}
Attention deficit with hyperactivity	{314.01}
Bipolar I disorder	{296.00, 296.01, 296.02, 296.03, 296.04, 296.05, 296.06, 296.40, 296.41, 296.42, 296.43, 296.44, 296.45, 296.46, 296.60, 296.61, 296.62, 296.63, 296.64, 296.65, 296.66, 296.50, 296.51, 296.52, 296.53, 296.54, 296.55, 296.56, 296.7}
Bladder cancer	{188.0, 188.1, 188.2, 188.3, 188.4, 188.5, 188.6, 188.7, 188.8, 188.9}
Breast cancer	{174.0, 174.1, 174.2, 174.3, 174.4, 174.5, 174.6, 174.8, 174.9, 175.0, 175.9}
Coronary disease	{402.0, 402.01, 402.10, 402.11, 402.90, 402.91, 403.00, 403.01, 403.10, 403.11, 403.90, 403.91, 404.00, 404.01, 404.02, 404.03, 404.10, 404.11, 404.12, 404.13, 404.90, 404.91, 404.92, 404.93, 405.01, 405.09, 405.11, 405.19, 405.91, 405.99, 410.00, 410.01, 410.02, 410.10, 410.11, 410.12,410.20, 410.21, 410.22, 410.30, 410.31, 410.32, 410.40, 410.41, 410.42, 410.50, 410.51, 410.52, 410.60, 410.61, 410.62, 410.70, 410.71, 410.72, 410.80, 410.81, 410.82, 410.90, 410.91, 410.92, 411.0, 411.1, 411.81, 411.89, 412, 413.0, 413.1, 413.9, 414.00, 414.01, 414.02, 414.03, 414.04, 414.05, 414.06, 414.07, 414.10, 414.11, 414.12, 414.19, 414.2, 414.3, 414.8, 414.9, 415.0, 415.11, 415.12, 415.19, 416.0, 416.1, 416.8, 416.9, 417.0, 417.1, 417.8, 417.9, 420.0, 420.90, 420.91, 420.99, 430, 431, 432.0, 432.1, 432.9, 433.0, 433.00, 433.01, 433.10, 433.11, 433.20, 433.21, 433.30, 433.31, 433.80, 433.81, 433.90, 433.91, 434.00, 434.01, 434.10, 434.11, 434.90, 434.91, 435.0, 435.1, 435.2, 435.3, 435.8, 435.9, 436, 437.0, 437.1, 437.2, 437.3, 437.4, 437.5, 437.6, 437.7, 437.8, 437.9, 438.10, 438.11, 438.12, 438.19, 438.20, 438.21, 438.22, 438.30, 438.31, 438.32, 438.40, 438.41, 438.42, 438.50, 438.51, 438.52, 438.53, 438.6, 438.7, 438.81, 438.82, 438.83, 483.84, 438.85, 438.89, 438.9, 440.0, 440.1, 440.2, 440.21, 440.22, 440.23, 440.24, 440.29, 440.30, 440.31, 440.32, 440.4, 440.8, 440.9, 441.00, 441.01, 441.02, 441.03, 441.1, 441.2, 441.3, 441.4, 441.5, 441.6, 441.7, 441.9, 442.0, 442.1, 442.2, 442.3, 442.81, 442.82, 442.83, 442.84, 442.89, 442.9, 443.0, 443.1, 443.2, 443.21, 443.22, 443.23, 443.24, 443.29, 443.81, 443.82, 443.89, 443.9, 444.0, 444.1, 444.21, 444.22, 444.81, 444.89, 444.9, 445.01, 445.02, 445.81, 445.89, 446.0, 446.1, 446.20, 446.21, 446.29, 446.3, 446.4, 446.5, 446.6, 446.7, 447.0, 447.1, 447.2, 447.3, 447.4, 447.5, 447.6, 447.8, 447.9, 448.0, 448.1, 448.9}
Dental caries	{521.00, 521.01, 521.02, 521.03, 521.04, 521.05, 521.06, 521.07, 521.08, 521.09}
Diabetes mellitus type 1	{250.01, 250.03, 250.11, 250.13, 250.21, 250.23, 250.31, 250.33, 250.41, 250.43, 250.51, 250.53, 250.61, 250.63, 250.71, 250.73, 250.81, 250.83, 250.91, 250.93}
Diabetes mellitus type 2	{250.00, 250.02, 250.10, 250.12, 250.20, 250.22, 250.30, 250.32, 250.40, 250.42, 250.50, 250.52, 250.60, 250.62, 250.70, 250.72, 250.80, 250.82, 250.90, 250.92}
Lung cancer	{162.0, 162.2, 162.3, 162.4, 162.5, 162.8, 162.9}
Major depressive disorder	{296.2, 296.3}
Pancreatic cancer	{157.0, 157.1, 157.2, 157.3, 157.4, 157.8, 157.9}
Platelet phenotypes	{287.1}
Preterm birth	{644.00, 644.03, 644.10, 644.13, 644.20, 644.21, 765.00, 765.01, 765.02, 765.03, 765.04, 765.05, 765.06, 765.07, 765.08, 765.09, 765.10, 765.11, 765.12, 765.13, 765.14, 765.15, 765.16, 765.17, 765.17, 765.18, 765.19, 765.20, 765.21, 765.22, 765.23, 765.24, 765.25, 765.26, 765.27, 765.28, 765.29}
Prostate cancer	{185}
Psoriasis	{696.0, 696.1, 696.2, 696.3, 696.4, 696.5, 696.8}
Renal cancer	{189.1}
Schizophrenia	{295.00, 295.01, 295.02, 295.03, 295.04, 295.05, 295.10, 295.11, 295.12, 295.13, 295.14, 295.15, 295.20, 295.21, 295.22, 295.23, 295.24, 295.25, 295.30, 295.31, 295.32, 295.33, 295.34, 295.35, 295.40, 295.41, 295.42, 295.43, 295.44, 295.45, 295.50, 295.51, 295.52, 295.53, 295.54, 295.55, 295.60, 295.61, 295.62, 295.63, 295.64, 295.65, 295.70, 295.71, 295.72, 295.73, 295.74, 295.75, 295.80, 295.81, 295.82, 295.83, 295.84, 295.85, 295.90, 295.91, 295.92, 295.93, 295.94, 295.95}
Sickle-cell disease	{282.60, 282.61, 282.62, 282.63, 282.64, 282.68, 282.69}

Table S3. Satisfied utility constraints for $k = 2$ and the single-visit case (✓ denotes that a utility constraint is satisfied)

	VNEC		$VNEC_{KC}$	
Disease	UGACLIP	ACLIP	UGACLIP	ACLIP
Asthma	✓		✓	✓
Attention deficit with hyperactivity	✓		✓	
Bipolar I disorder	✓		✓	✓
Bladder cancer	✓		✓	✓
Breast cancer	✓		✓	
Coronary disease	✓		✓	✓
Dental caries	✓		✓	
Diabetes mellitus type 1	✓		✓	✓
Diabetes mellitus type 2	✓		✓	✓
Lung cancer	✓		✓	
Major depressive disorder			✓	
Pancreatic cancer	✓		✓	✓
Platelet phenotypes			✓	
Preterm birth	✓		✓	✓
Prostate cancer	✓		✓	
Psoriasis	✓		✓	✓
Renal cancer	✓		✓	
Schizophrenia			✓	
Sickle-cell disease			✓	

Table S4. Satisfied utility constraints for $k = 10$ and the single-visit case (✓ denotes that a utility constraint is satisfied)

	VNEC		$VNEC_{KC}$	
Disease	UGACLIP	ACLIP	UGACLIP	ACLIP
Asthma	✓		✓	
Attention deficit with hyperactivity				
Bipolar I disorder			✓	
Bladder cancer				
Breast cancer	✓		✓	
Coronary disease	✓		✓	✓
Dental caries	✓		✓	
Diabetes mellitus type 1	✓		✓	
Diabetes mellitus type 2	✓		✓	✓
Lung cancer				
Pancreatic cancer				
Platelet phenotypes				
Preterm birth			✓	
Prostate cancer			✓	
Psoriasis			✓	
Renal cancer			✓	
Schizophrenia			✓	

Table S5. Satisfied utility constraints for $k = 25$ and the single-visit case (✓ denotes that a utility constraint is satisfied)

	VNEC		$VNEC_{KC}$	
Disease	UGACLIP	ACLIP	UGACLIP	ACLIP
Asthma			✓	
Bipolar I disorder			✓	
Bladder cancer				
Breast cancer			✓	
Coronary disease	✓		✓	✓
Dental caries			✓	
Diabetes mellitus type 1	✓		✓	
Diabetes mellitus type 2	✓		✓	✓
Lung cancer				
Pancreatic cancer				
Platelet phenotypes				
Preterm birth			✓	
Prostate cancer				
Schizophrenia				

Table S6. Satisfied utility constraints for $k = 2$ and the all-visits case (✓ denotes that a utility constraint is satisfied)

	VNEC		$VNEC_{KC}$	
Disease	UGACLIP	ACLIP	UGACLIP	ACLIP
Asthma				
Attention deficit with hyperactivity				
Bipolar I disorder			✓	
Bladder cancer				
Breast cancer			✓	
Coronary disease				
Dental caries				
Diabetes mellitus type 1		✓	✓	
Diabetes mellitus type 2		✓	✓	
Lung cancer		✓	✓	
Pancreatic cancer		✓	✓	
Platelet phenotypes				
Preterm birth				
Prostate cancer		✓	✓	
Psoriasis				
Renal cancer				
Schizophrenia				
Sickle-cell disease				

Table S7. Satisfied utility constraints for $k = 5$ and the all-visits case (✓denotes that a utility constraint is satisfied)

	VNEC		$VNEC_{KC}$	
Disease	UGACLIP	ACLIP	UGACLIP	ACLIP
Asthma				
Attention deficit with hyperactivity				
Bipolar I disorder				
Bladder cancer				
Breast cancer				
Coronary disease				
Dental caries				
Diabetes mellitus type 1	✓		✓	
Diabetes mellitus type 2	✓		✓	
Lung cancer	✓		✓	
Pancreatic cancer				
Platelet phenotypes				
Preterm birth				
Prostate cancer				
Psoriasis				
Renal cancer				
Schizophrenia				
Sickle-cell disease				

Table S8. Satisfied utility constraints for $k = 10$ and the all-visits case (✓ denotes that a utility constraint is satisfied)

	VNEC		$VNEC_{KC}$	
Disease	UGACLIP	ACLIP	UGACLIP	ACLIP
Asthma				
Attention deficit with hyperactivity				
Bipolar I disorder				
Bladder cancer				
Breast cancer				
Coronary disease				
Dental caries				
Diabetes mellitus type 1	✓		✓	
Diabetes mellitus type 2	✓		✓	
Lung cancer	✓		✓	
Pancreatic cancer				
Platelet phenotypes				
Preterm birth				
Prostate cancer				
Psoriasis				
Renal cancer				
Schizophrenia				

Table S9. Satisfied utility constraints for $k = 25$ and the all-visits case (✓ denotes that a utility constraint is satisfied)

	VNEC		$VNEC_{KC}$	
Disease	UGACLIP	ACLIP	UGACLIP	ACLIP
Asthma				
Bipolar I disorder				
Bladder cancer				
Breast cancer				
Coronary disease				
Dental caries				
Diabetes mellitus type 1	✓		✓	
Diabetes mellitus type 2	✓		✓	
Lung cancer	✓		✓	
Pancreatic cancer				
Platelet phenotypes				
Preterm birth				
Prostate cancer				
Schizophrenia				