Identifying Most Lethal Cliques in Disease Comorbidity Graphs

Parisa Vaghfi Mohebbi[a], Yajun Lu[b], Zhuqi Miao[c], Balabhaskar Balasundaram[a], Pankush Kalgotra[d], and Ramesh Sharda[e]

[a]School of Industrial Engineering & Management, Oklahoma State University, Stillwater, Oklahoma, United States; [b]Department of Management & Marketing, Jacksonville State University, Jacksonville, Alabama, United States; [c]School of Business, State University of New York at New Paltz, New Paltz, New York, United States; [d] Department of Business Analytics and Information Systems, Auburn University, Auburn, Alabama, United States; [e]Department of Management Science and Information Systems, Oklahoma State University, Stillwater, Oklahoma, United States

ARTICLE HISTORY
Compiled February 20, 2026

ABSTRACT
Mortality rate refers to the overall likelihood of death within a specific population over a defined period. The knowledge of high mortality rate disease clusters can enable healthcare providers and patients to be proactive and develop tailored interventions that improve patient outcomes. In this paper, we introduce a methodology for systematically incorporating electronic health record data in an optimization framework to find cliques of comorbid diseases that correspond to the highest mortality rates among a given patient population. To this end, we introduce *the maximum mortality rate clique problem* and devise two approaches to solve it: (*i*) we formulate a mathematical optimization model that maximizes a fractional objective function subject to linear constraints in binary variables, which we reformulate as a mixed-integer linear program and solve using a commercial solver with delayed constraint generation, and (*ii*) we design an enumerative combinatorial algorithm based on the classical Bron–Kerbosch algorithm for enumerating maximal cliques. We conduct a detailed computational study and report results from our experiments with both approaches on a test bed of instances derived from millions of de-identified patient electronic health records.

KEYWORDS
Clique enumeration; Bron–Kerbosch algorithm; Fractional programming; Mortality rate; Comorbidity graph; Electronic health record analysis

1. Introduction

Mortality rate, the fraction of deceased patients who had a particular disease or group of diseases, is a metric of broad interest in medicine, epidemiology, and public health. Analyzing the mortality rates among patients with multiple illnesses enables healthcare providers to devise effective strategies for patient care. The notion commonly used to discuss the presence of multiple diseases in a single patient is known as *comorbidity*. Comorbidity assumes particular significance in the context of mortality, as the interaction between two or more diseases can result in health outcomes that differ from those produced by each disease on its own (Feinstein, 1970). Individuals expe-

CONTACT Balabhaskar Balasundaram. Email: baski@okstate.edu

riencing multiple health issues often confront an elevated mortality risk attributed to the interplay among their conditions. Therefore, it is imperative to take comorbidities into account when estimating mortality rates.

Gijsen et al. (2001) conducted a comprehensive review of existing literature to identify causes and consequences of comorbidity, with a strong emphasis on how comorbidity influences mortality rates. Comorbidity was consistently associated with increased mortality rates, underscoring its impact on patient outcomes. Previous research indicates that patients who have been diagnosed with diseases such as myocardial infarction, diabetes, chronic obstructive pulmonary disease, chronic kidney disease, dementia, depression, hip fracture, stroke, colorectal cancer, and lung cancer, are at an elevated risk of developing heart failure, which can ultimately lead to fatal outcomes (Ahluwalia et al., 2012). This connection between various diseases and heart failure is particularly noteworthy as a majority of reported cases in the United States involve heart-related diseases and cancer (Xu, Murphy, Kochanek, & Arias, 2016). The importance of detecting comorbid diseases is underscored by Braunstein et al. (2003) in their study which highlights that elderly patients with chronic heart failure often succumb to non-cardiac conditions. Their research suggests that identifying and addressing these non-cardiac conditions in heart patients can lead to improved health outcomes. In a population-based cohort study undertaken by Corraini et al. (2018), the impact of comorbidity on post-stroke mortality rates was investigated. Their findings showed that comorbidity, particularly conditions such as cancer, advanced renal disease, or liver disease, significantly increased one-year mortality rates after a stroke. Additionally, Redelmeier, Tan, and Booth (1998) argue that the presence of a critical disease in a patient can monopolize healthcare attention, potentially leading to the neglect of unrelated diseases. Hence, taking comorbidities into account can significantly enhance a patient's overall condition and, in turn, contribute to the overall improvement of the healthcare system.

Identifying comorbidities in the form of disease clusters with a high mortality rate can have several benefits in epidemiology, health policy, healthcare management, and public health planning to monitor and analyze trends in health and disease. With the knowledge of such disease clusters, healthcare providers can anticipate the onset of diseases early and intervene proactively, significantly enhancing the likelihood of successful treatment and lowering mortality rates. Efficient resource allocation becomes possible as health system administrators can direct resources towards treating diseases more likely to lead to death, ultimately improving overall patient outcomes.

Moreover, the identification of high mortality rate clusters can serve as a catalyst for further research into their joint impact, the underlying causes of these diseases, and potential treatments, thereby fostering the development of new therapies and improving patient well-being. Public health officials can leverage this information to identify high-risk geographic areas or populations, and design targeted interventions aimed at reducing mortality rates within those communities, thereby promoting overall public health.

Although comorbidities have been studied before in specific clinical contexts, availability of large-scale electronic health records affords the opportunity to identify a variety of such disease clusters that have led to increased mortality rates across the full spectrum of diseases. This multifaceted approach, combining clinical insights and public health strategies, represents a potent tool for improving healthcare and reducing mortality on multiple fronts.

Our main contributions in this article address a gap in the healthcare analytics literature by developing an optimization-based analytical framework for the problem of

finding clusters of comorbid diseases with high mortality rates. We introduce a unified framework for analyzing large-scale electronic health records to find lethal comorbidities by framing the question of interest as a combinatorial optimization problem. We then develop two exact algorithmic approaches for solving the problem that are evaluated in a computational study.

The remainder of this paper is structured as follows. In Section 2, we provide a brief review of prior work related to mortality rates and comorbidity analysis, followed by a formal statement of the optimization problem of interest in Section 3. In Section 4, we present two methodologies for solving the optimization problem of interest. We describe the process used to develop the dataset used in our study in Section 5. Then, we assess the effectiveness of our proposed approaches, examine their scalability, and identify the most lethal cliques in our dataset in Section 6. We conclude in Section 7 with a discussion of our contributions, the implications for healthcare research, limitations of our study, and identify some directions for future research.

2. Related work

Analytical approaches leveraging comorbidities in studying mortality rates are predominantly statistical in nature. In this section, we briefly review some statistical and empirical studies that focused on the impact of comorbidities on mortality rates.

Zolbanin, Delen, and Zadeh (2015) demonstrate that considering comorbidity will improve the prediction performance of the models developed to forecast the survivability rate of patients with cancer. Various related studies have explored comorbidities with a particular emphasis on gender-related factors (Short, Yang, & Jenkins, 2013). A majority of studies in the literature primarily aim to identify comorbidities linked to specific conditions like diabetes, obesity, and cancer, among others (Holguin, Folch, Redd, & Mannino, 2005; Leontiadis, Molloy-Bland, Moayyedi, & Howden, 2013; Marrie et al., 2015). For example, Van Gestel et al. (2013) investigated the impact of comorbidity and age on postoperative mortality in gastrointestinal cancer patients. Their study found that comorbidity and older age were associated with early postoperative mortality after cancer resection. The presence of specific comorbidities such as cardiovascular diseases had a significant impact on 30-day mortality rates.

Koskinen et al. (2022) employed a data-driven approach to analyze comorbidity patterns in a wide range of common diseases. Their study identified disease-specific patient subgroups with distinctive diagnosis patterns, survival functions, and laboratory correlates. This research emphasized the heterogeneity within disease populations and the need for personalized care and risk assessment. W. S. Lu and Tsutakawa (1996) proposed a mixed Poisson regression model for analyzing mortality data, placing particular emphasis on understanding age-specific and age-standardized mortality rates. Their approach utilized a marginal quasi-likelihood function to estimate mortality rates in an empirical Bayes manner. This method aimed to improve the practicality of analyzing mortality data, emphasizing marginal mortality rates for diseases like lung cancer. Y. Lu, Chen, Miao, Delen, and Gin (2021) modeled comorbidities as temporal disease networks to visualize comorbidity progression and to identify critical points in the progression timeline using clustering techniques.

The literature contains plenty of medical, statistical, and empirical evidence establishing the impact of comorbidities on mortality rates. However, we are unable to find optimization-based decision-support tools capable of dealing with a massive database of electronic health records, which can expand the toolkit of healthcare prac-

titioners. This is where our most significant contributions lie. We introduce the optimization problem of interest formally in the next section.

3. Problem statement

We formalize as an optimization problem, the problem of identifying disease clusters that correspond to high mortality rates among patients represented by the electronic health record (EHR) dataset under consideration. It is not necessary that this dataset represent the general population and we can filter patients in or out based on additional criteria like age groups, demographics, geographical location, presence or absence of a particular disease, or any combination thereof. We assume EHR data is available that contains the relevant information and lay the groundwork for our problem statement next.

3.1. Notations and definitions

We denote the set of patients encountered within the time frame covered by the EHRs as P, and those who died within this time frame are represented by $\widetilde{P} \subseteq P$. Let V denote the set of diseases diagnosed in the patients in P. We assume both sets are minimal in the sense that each patient in the set P is affected by at least one disease from the set V and each disease from set V afflicts at least one patient in P. To capture the incidence of diseases in patients, we use disease-patient incidence sets, where $A_u \subseteq P$ represents the subset of patients afflicted with disease $u \in V$. For convenience, we also use a patient-disease incidence set, where $D_i \subseteq V$ denotes the subset of diseases afflicting patient $i \in P$.

We formalize the notion of disease comorbidity, as it has been in some prior works, using a graph representation (Hidalgo, Blumm, Barabási, & Christakis, 2009; Kalgotra, Sharda, & Croff, 2017; Y. Lu et al., 2021). Given an EHR dataset documenting patient encounters over a specified time frame, we can construct a *comorbidity graph* $G = (V, E)$ in which the vertex set V represents the set of diseases and the edge set E contains an edge $\{u, v\}$ if and only if diseases u and v have frequent co-occurrence among patients in the EHR data.

A *clique*, which is a subset of pairwise adjacent vertices, is used to model a cluster of comorbid diseases in this study. Given a non-empty clique of diseases $C \subseteq V$, its mortality rate $\mu(C)$ can be defined as follows (Porta, 2014):

$$\mu(C) = \frac{\left| \widetilde{P} \cap \bigcap\limits_{u \in C} A_u \right|}{\left| \bigcap\limits_{u \in C} A_u \right|}, \tag{1}$$

where $\mu(C)$ is taken to be zero if the denominator vanishes. Implicit in this definition is that $\mu(C)$ is associated with a particular length of time over which the EHRs are acquired. We remark that this definition of mortality rate is a suitable metric for our decision-support problem, but variants that focus on specific age groups (e.g., infant), geographical locations, conditions (e.g., maternal), diseases (e.g., COVID-19), time periods (e.g., 1-year, 30-day) are also widely used in other studies (Baud et al., 2020; Chen, Normand, Wang, & Krumholz, 2011; Marshall, 2018; Osmond, 1985).

3.2. The maximum mortality rate clique problem

Given positive integers b, ℓ, our problem of interest is to find a clique C in G with at most b diseases that at least ℓ patients in the data set are simultaneously afflicted by, for which the mortality rate $\mu(C)$ is a maximum. That is, we aim to solve the following optimization problem:

$$\max \left\{ \mu(C) : C \text{ is a clique in } G, |C| \leqslant b, \left| \bigcap_{u \in C} A_u \right| \geqslant \ell \right\}. \tag{2}$$

The upper-bound b on clique size ensures that we identify smaller clusters associated with elevated mortality rates, as they hold greater diagnostic significance. It is usually the case that the mortality rates are higher for patients that have a large number of comorbidities. The lower-bound ℓ on the number of patients simultaneously afflicted with all diseases in C ensures that the calculated mortality rate $\mu(C)$ is a reliable and indicative estimate from this dataset.

A variant of problem (2) is to identify a maximum *marginal* mortality rate clique. Given a clique of diseases, we are interested in identifying diseases, the addition of which maximizes the overall mortality rate of the new clique. This further enhances the prognosis potential of this methodology for patient populations already afflicted with some diseases. Specifically, given a pre-established clique of strictly fewer than b diseases denoted by C^0, we aim to identify at most $b - |C^0|$ additional disease(s) that maintain the clique property when added to C^0 and maximize the mortality rate of the new clique. That is, we now aim to solve the following optimization problem:

$$\max \left\{ \mu(C) : C \text{ is a clique in } G, C \supseteq C^0, |C| \leqslant b, \left| \bigcap_{u \in C} A_u \right| \geqslant \ell \right\}. \tag{3}$$

It is easy to see that the maximum marginal mortality rate clique problem (3) includes optimization problem (2) as a special case, obtained when $C^0 = \varnothing$. Next, we introduce two methods for exactly solving the maximum mortality rate clique problem, which can be easily extended to solve the marginal mortality rate counterpart.

4. Methodology

At the outset, we must recognize that the mortality rate function $\mu : 2^V \longrightarrow \mathbb{R}^+$ is neither additive nor monotonic. For some $v \in V \backslash C$,

$$\mu(C \cup \{v\}) = \frac{\left| \widetilde{P} \cap \bigcap\limits_{u \in C} A_u \cap A_v \right|}{\left| \bigcap\limits_{u \in C} A_u \cap A_v \right|}, \tag{4}$$

implying that $\mu(C \cup \{v\})$ is not separable into $\mu(C) + g(\{v\})$ for some function $g(.)$ that depends on v (and is independent of C). The addition of a new disease v to C decreases the numerator and denominator of $\mu(C)$ by potentially different amounts. Although the decrease in the numerator is no larger than the decrease in the denominator, $\mu(.)$ is not monotonic as $\mu(C \cup \{v\})$ and $\mu(C)$ are incomparable as we elaborate next.

Trivially, $\mu(C \cup \{v\}) = \mu(C)$ if $A_v \supseteq \bigcap_{u \in C} A_u$. Suppose that is not the case and $\bigcap_{u \in C} A_u \backslash A_v$ is nonempty. Let κ_a and κ_e correspond to the number of patients alive and deceased, respectively, of the total number of patients that have all the diseases in C but not disease v, i.e., $\kappa_a + \kappa_e = \left| \bigcap_{u \in C} A_u \backslash A_v \right|$. We can rewrite equation (4) as,

$$\mu(C \cup \{v\}) = \frac{\left| \widetilde{P} \cap \bigcap_{u \in C} A_u \right| - \kappa_e}{\left| \bigcap_{u \in C} A_u \right| - \kappa_a - \kappa_e}.$$

It is easy to see that if $\kappa_a = 1$ and $\kappa_e = 0$, then $\mu(C \cup \{v\}) > \mu(C)$. Alternately, if $\kappa_a = 0$ and $\kappa_e = 1$, then assuming $\left| \bigcap_{u \in C} A_u \right| - \kappa_e \geqslant 1$ and $\mu(C) > 0$, we can deduce that $\mu(C \cup \{v\}) \leqslant \mu(C)$ as $\mu(C) \leqslant 1$.

This characteristic of the mortality rate function is an impediment to the development of exact combinatorial branch-and-bound or dynamic programming algorithms that usually exploit such properties. A mixed-integer linear program (MILP) is therefore a natural choice to model and solve problem (3). Furthermore, as we are specifically interested in cliques of small sizes with high mortality rates, an enumerative algorithm could potentially be a practical choice. Hence, we approach the task of solving combinatorial optimization problems (2) and (3) to identify a maximum (marginal) mortality rate clique in the comorbidity graph G using the following two methodologies: the first involves modeling the problem as a mathematical optimization formulation that maximizes a single fractional objective function subject to linear constraints in binary variables, then linearizing it to an MILP and developing decomposition techniques that can be implemented using a general purpose MILP solver, while the other approach is to design a recursive algorithm that enumerates cliques of size at most b.

4.1. An MILP formulation and delayed constraint generation

The maximum mortality rate clique problem (2) is formulated in (5). We use the binary decision variable $x_j = 1$ to indicate that disease $j \in V$ is selected in the clique; $x_j = 0$ otherwise. For each patient $i \in P$, we can associate a binary variable $y_i = 1$ to indicate that patient i has all diseases selected in the clique $C = \{j \in V : x_j = 1\}$. Henceforth, we use $\bar{G} = (V, \bar{E})$ to denote the complement graph of G. We wish to maximize the mortality rate in the fractional objective function (5a). Constraint (5b) ensures that the subset of selected diseases forms a clique in G by preventing non-adjacent pairs of vertices from being included simultaneously. The size of the selected clique is bounded from above by constraint (5c). Constraint (5d) ensures that if a disease j not afflicting patient i is included in the clique, the y_i is forced to zero. Or equivalently, if $y_i = 1$, then $x_j = 0$ for every $j \in V \backslash D_i$. The converse is enforced by constraint (5e) as $y_i = 0$ forces the inclusion of at least one disease not afflicting patient i in the clique. Or equivalently, if $x_j = 0$ for every $j \in V \backslash D_i$, then $y_i = 1$. Together, constraints (5d) and (5e) ensure that $y_i = 1$ if and only if patient i is afflicted with every disease included in the selected clique represented by the incidence vector x. In order for the mortality rate to be indicative and to avoid trivial solutions, constraint (5f) imposes

a lower bound ℓ on the number of patients that have all of the diseases included in the clique. For solving the marginal mortality rate clique problem (3), the diseases in clique C^0 can be forced to be included by adding the constraints $x_j = 1$ for each $j \in C^0$.

$$\max \frac{\sum_{i \in \tilde{P}} y_i}{\sum_{i \in P} y_i} \tag{5a}$$

$$s.t. \quad x_u + x_v \leqslant 1 \qquad \forall \{u, v\} \in \bar{E} \tag{5b}$$

$$\sum_{j \in V} x_j \leqslant b \tag{5c}$$

$$y_i \leqslant 1 - x_j \qquad \forall j \in V \backslash D_i, i \in P \tag{5d}$$

$$y_i \geqslant 1 - \sum_{j \in V \backslash D_i} x_j \qquad \forall i \in P \tag{5e}$$

$$\sum_{i \in P} y_i \geqslant \ell \tag{5f}$$

$$x_j \in \{0, 1\} \qquad \forall j \in V \tag{5g}$$

$$y_i \in \{0, 1\} \qquad \forall i \in P \tag{5h}$$

Formulation (6) that follows is an MILP that is equivalent to the fractional program (5) obtained by linearizing the fractional objective function. Objective function (5a) is substituted using a new variable as $z = \sum_{i \in \tilde{P}} y_i / \sum_{i \in P} y_i$, and we introduce new continuous variables $w_i \in [0, 1]$ for each $i \in P$ to linearize the product zy_i. If $y_i = 0$, constraint (6b) forces $w_i = 0$, while constraints (6c) and (6d) are redundant. If $y_i = 1$, constraints (6c) and (6d) force $w_i = z$, with constraint (6b) being redundant. Effectively, constraints (6b)–(6d) force each w_i to equal zy_i, and consequently, equation (6e) models the objective function z as desired.

In addition to linearization, we make two further changes to formulation (5) based on preliminary computational experiments as explained next. Formulation (6) reduces the number of conflict constraints (5b) by using binary variables that indicate the component that contains the clique. We let $\mathcal{C}$ denote the collection of connected components of G and associate a binary variable f_H with each connected component $H \in \mathcal{C}$. For each component $H \in \mathcal{C}$, we denote its vertex set by $V(H)$, the edge set by $E(H)$, and the complement graph by $\bar{H}$. Conflict constraints (6f) are written only for non-adjacent pairs of vertices inside the connected component, while constraints (6g) and (6h) force the x variables corresponding to all but one connected component to zero. The other change we make is to replace constraints (5d) with its aggregated counterpart (6j), which led to a noticeable reduction in overall running times in our preliminary experiments, possibly due to the reduced size of the basis when solving the node linear programming (LP) relaxations.

In our preliminary experiments, solving formulation (6) directly using the solver was not viable for instances with a large number of patients. Consequently, we introduce a decomposition branch-and-cut (DBC) approach with delayed generation of constraints (6b)–(6d), which grow with the number of patients in the dataset. We also experimented with the delayed generation of constraints (6j)–(6k) in combination with delayed generation of constraints (6b)–(6d), and by themselves, as this group

of constraints also grows with $|P|$. However, delayed generation of constraints (6b)–(6d) produced the best overall performance for the DBC approach in our preliminary experiments.

$$\max z \tag{6a}$$

$$s.t. \quad w_i \leqslant y_i \qquad \forall i \in P \tag{6b}$$

$$w_i \geqslant z + y_i - 1 \qquad \forall i \in P \tag{6c}$$

$$w_i \leqslant z \qquad \forall i \in P \tag{6d}$$

$$\sum_{i \in \widetilde{P}} y_i = \sum_{i \in P} w_i \tag{6e}$$

$$x_u + x_v \leqslant 1 \qquad \forall \{u, v\} \in E(\bar{H}), H \in \mathcal{C} \tag{6f}$$

$$\sum_{H \in \mathcal{C}} f_H = 1 \tag{6g}$$

$$x_j \leqslant f_H \qquad \forall j \in V(H), H \in \mathcal{C} \tag{6h}$$

$$\sum_{j \in V} x_j \leqslant b \tag{6i}$$

$$|V \backslash D_i| y_i \leqslant |V \backslash D_i| - \sum_{j \in V \backslash D_i} x_j \qquad \forall i \in P \tag{6j}$$

$$y_i \geqslant 1 - \sum_{j \in V \backslash D_i} x_j \qquad \forall i \in P \tag{6k}$$

$$\sum_{i \in P} y_i \geqslant \ell \tag{6l}$$

$$z \in [0, 1] \tag{6m}$$

$$w_i \in [0, 1] \qquad \forall i \in P \tag{6n}$$

$$x_j \in \{0, 1\} \qquad \forall j \in V \tag{6o}$$

$$y_i \in \{0, 1\} \qquad \forall i \in P \tag{6p}$$

$$f_H \in \{0, 1\} \qquad \forall H \in \mathcal{C} \tag{6q}$$

4.2. An enumerative approach based on the Bron–Kerbosch algorithm

The original Bron–Kerbosch (BK) algorithm is a recursive algorithm that operates with three sets: a current clique C, set L containing every candidate vertex that may be added to enlarge clique C, and set S containing vertices that are adjacent to all vertices in C but excluded as candidates for addition, in order to avoid outputting the same maximal clique multiple times during the recursive calls. The algorithm selects a vertex from L and moves it to C. After updating L and S by removing the non-neighbors of the recently added vertex v, the recursive function is called with the new C, L, and S. When L and S are empty, the current clique is returned as it is a maximal clique that has not been previously output. Then, the algorithm backtracks to the most recent recursive call and the candidate vertex that was added is removed from the set of candidates L and added to S to help track for duplicates.

Several algorithmic variants for enumerating cliques in a graph have been introduced since the classical algorithm by Bron and Kerbosch (1973); see for instance (Cazals & Karande, 2008; Chiba & Nishizeki, 1985; Eppstein, Löffler, & Strash, 2013; San Segundo, Artieda, & Strash, 2018; Tomita, Tanaka, & Takahashi, 2006). Re-

call our discussion in Section 4 on the mortality rate function, which is not monotonic. Therefore, we cannot restrict our search only to cliques that are maximal with respect to the inclusion of vertices. Consequently, many of the enhancements to the classical algorithm (e.g., pivoting) are not applicable for our problem. The core of our algorithm is a recursion that is similar to the original algorithm in terms of enumerating cliques. However, as the mortality rate function is not monotonic, we must consider *all cliques*, not just maximal cliques.

Algorithm 1: Enumerating Highest (Marginal) Mortality Rate Cliques

Input: $\langle P, \widetilde{P}, A, G = (V, E), \ell, b \rangle$ and initial clique C^0 (possibly empty) such that $|C^0| < b$

Output: Max-heap Q of feasible cliques with mortality rate as priority

$Q \leftarrow \varnothing$
if $C^0 \neq \varnothing$ **then**
  Insert C^0 in Q with priority $\mu(C^0)$
  $C \leftarrow C^0$, $L \leftarrow \bigcap_{u \in C^0} N(u)$, and $den \leftarrow \bigcap_{u \in C^0} A_u$
else
  $C \leftarrow \varnothing$, $L \leftarrow V$, and $den \leftarrow P$
`Bron-Kerbosch`(C, L, den)
return Q
Function `Bron-Kerbosch`(C, L, den)
  if $|C| = b$ *or* $L = \varnothing$ **then**
    return
  for $v \in L$ **do**
    $newden \leftarrow den \cap A_v$
    if $|newden| \geqslant \ell$ **then**
      $\mu \leftarrow \frac{|\widetilde{P} \cap newden|}{|newden|}$
      $C \leftarrow C \cup \{v\}$
      if $C \notin Q$ **then**
        Insert C in Q with priority μ
      `Bron-Kerbosch`$(C, L \cap N(v), newden)$
    $L \leftarrow L \backslash \{v\}$
  return

We introduce Algorithm 1 to solve both problems (2) and (3), which includes pruning steps based on the following feasibility conditions: *(i)* restricting the enumeration to cliques of size at most b, and *(ii)* eliminating cliques associated with insufficient patients by enforcing that $\left| \bigcap_{u \in C} A_u \right| \geqslant \ell$. In Algorithm 1, we use $N(v) := \{u : \{u, v\} \in E\}$ to denote the neighbors of vertex v in G.

A global max-heap Q, prioritized by mortality rate, is used to store the enumerated cliques and if non-empty, C^0 is inserted in Q initially. The disjoint sets C and L along with the set *den*, which represents the set $\bigcap_{u \in C} A_u$ from the denominator of the mortality rate function are initialized next. Due to the recursive nature of the algorithm, it is more efficient to track and incrementally update *den*, rather than re-

compute the intersection of several sets every time. The current clique C is initialized to C^0, the candidate set L is initialized to the common neighbors of all the vertices in C^0 (or the entire vertex set if C^0 is empty), and *den* is initialized to the set of patients that have all diseases in C^0 (or the entire patient set if C^0 is empty).

Our algorithm terminates the recursion if one of the following conditions is met in Step 10 of Algorithm 1: *(i)* when L is empty, which means that the current clique is maximal; or *(ii)* the size of current clique is b. As a clique is enlarged with the addition of a vertex v in a recursive call, the set *den* is updated, lower bound condition checked and if met, mortality rate is calculated. The new clique is inserted in Q with its mortality rate as priority if it is not already in the max-heap. Then a deeper recursive call is made with the updated sets. If candidate vertex $v \in L$ cannot be included due to the lower-bound ℓ not being satisfied, or if we added vertex v and completed that recursive call, vertex v is removed from the set L. After the recursive calls are exhausted, the max-heap Q contains all feasible cliques encountered, with their associated mortality rate as heap priority. We can now extract the clique with the highest (marginal) mortality rate and solve problems (2) and (3). More importantly, we can now return for any positive integer K, the cliques with the K highest mortality rates.

5. Preparing the dataset

Our computational study in Section 6 is based on a real, large-scale, de-identified EHR database from which we developed the instances in our test bed. An EHR is an electronic version of the medical history of patients that is maintained by a provider over time and includes all the key administrative clinical data relevant to a person's care such as demographics, diagnoses, procedures, progress notes, medications, vital signs, past medical history, immunizations, and laboratory data.

We evaluate our methodologies on a test bed derived from the Cerner Health Facts® EHR warehouse containing health records of 69 million unique patients that were contributed voluntarily across 200 hospitals in the United States (1999–2018) operating with Cerner systems (now acquired by Oracle Corp.). The EHR warehouse has been completely de-identified in compliance with Health Insurance Portability and Accountability Act (HIPAA) regulations.

5.1. Processing the EHR database

Our EHR dataset captures longitudinal data that tracks patient encounters over time, providing a continuous record of patient progress. For our purpose, the mortality status is captured using discharge information, and included both in-hospital mortality (expiration within the hospital) and hospice enrollment. Recall that the main inputs to our optimization problems are the set of patients (and the subset of deceased patients), disease–patience incidence sets, the comorbidity graph, and the user-specified bounds. Figure 1 describes the process we followed to create our dataset from the EHR database and includes associated details on the number of patients and encounters in the dataset.

In broad terms, we can describe our approach and our rationale behind the steps in Figure 1 as follows. First, we filter the database by the age group of the patients, restricting our attention to a five-year retrospective time window. Life expectancy at birth for the U.S. population in the time period corresponding to the EHRs is estimated between 76.7 and 78.7 years of age (Arias & Xu, 2022). To avoid skewing

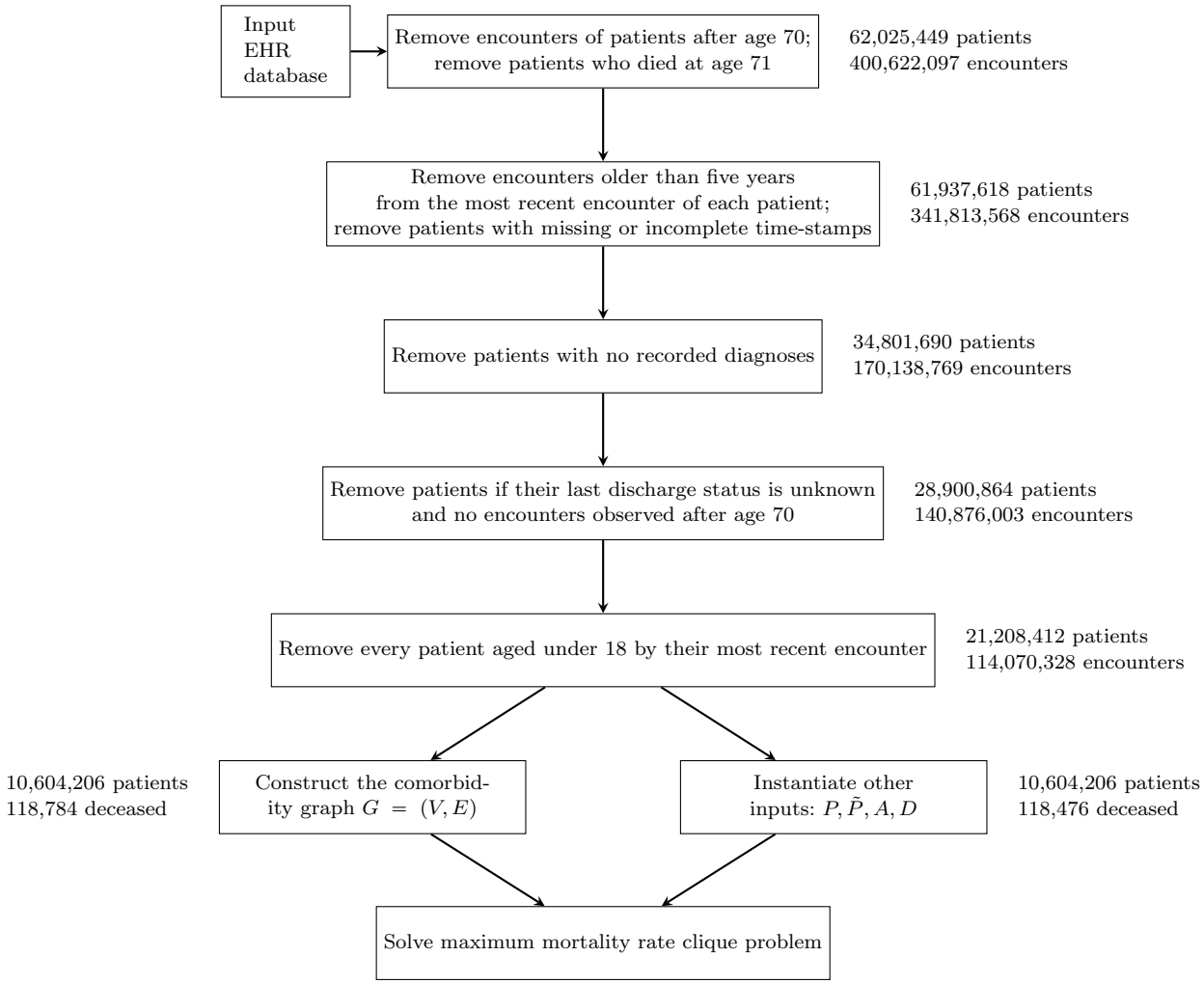

Figure 1. The process followed in deriving our case study dataset from the Cerner Health Facts EHR database. Beside each box we report the number of patients and encounters remaining in the dataset after executing the steps inside each box.

the mortality rate due to natural mortality in older patients, as our first step, we restrict our analysis to encounters of patients up to 70 years of age at the time of the encounter. Moreover, we follow the patient records for one more year and exclude those patients (along with all their encounters) who died at age 71 to remove the bias introduced by including the patients in the dataset who died immediately after the cut-off age of 70 years. We only include patients at least 18 years of age by their most recent encounter in our study, as the treatment paths for pediatrics are different. It is not uncommon to use age limits and time windows in comorbidity analysis. For instance, in their study, Hidalgo et al. (2009) consider patients aged 65 or older at the time of their first recorded encounter in their dataset and study patient mortality in eight years since the first diagnosis in their dataset. For each patient, we only include encounters from the most recent five years, in order to focus on the recent medical history of each patient.

Our next major step addressed missing data elements, specifically, removing patients with no recorded diagnoses, missing or incomplete time-stamps of encounters, and those with last discharge status unknown. However, there is no evidence to sus-

pect that this could introduce systematic bias, especially considering the massive size of the final dataset. Encounters may lack diagnosis codes due to coding errors, privacy considerations, and involvement limited to ancillary services without clinical evaluation, among other factors. Similarly, time-stamps could be missing due to data entry errors. The discharge status of each EHR encounter records whether the patient died. If the most recent discharge status of a particular patient is not known, it is difficult to discern the actual status of the patient. Notably, there were patients with unknown last discharge status in the dataset with encounters up to 70 years of age but visited at least once after 70 years of age. Therefore, we only removed a patient in our age group from the records if the last discharge status is unknown and no information about them was available after age 70.

After completing the foregoing steps the dataset contained approximately 21.2 million patients records, which was then divided into two random samples of equal size (i.e., approximately 10.6 million patients each). The first partition was used to create the comorbidity graph and the second partition was used to create the remaining inputs to our optimization problems. The second partition contains 118,476 patients that died or were assigned a status of hospice, corresponding to an overall mortality rate of 1.12%.

5.2. Constructing the comorbidity graph

The diagnoses in our EHR dataset are recorded according to the International Classification of Diseases 9^{th} and 10^{th} revision Clinical Modification (ICD-9-CM and ICD-10-CM) (Centers for Disease Control and Prevention, 2013, 2022). For our study we only need the disease categories, IDs of patients, and their mortality status. Given the presence of both ICD-9 and ICD-10 codes in our EHR data, the same disease may be represented by two different coding systems, introducing duplicates into our analysis. Furthermore, the extensive granularity of ICD codes can exacerbate this duplication. Specifically, ICD systems allocate distinct codes for diseases that, from a practical standpoint, are considered the same or very similar but may exhibit slight differences, such as variations in body locations (e.g., Osteoarthritis of the knee: M17.9 compared to that of the hip: M16.9), different contexts (e.g., fall from an escalator: W10.0 versus from a sidewalk curb: W10.1), and various causes (e.g., anthrax septicemia: 022.3 versus salmonella septicemia: 003.1). Therefore, we employed the Clinical Classifications Software (CCS) developed by Agency for Healthcare Research and Quality (AHRQ) (2021) to aggregate ICD-9/10 codes, grouping them into relatively high-level disease states for analysis. CCS, originally developed for classifying ICD-9 codes, has since been expanded into CCS Refined (CCSR) for ICD-10. This system has been widely applied in health data analytical research to group ICD codes into medically meaningful categories, thereby simplifying the classification space and enhancing the interpretability of analytical results (Kansal et al., 2021; Lee, Levin, Finley, & Heilig, 2019; Malecki et al., 2024).

Each high-level disease associated with the 10.6 million patients is represented as a vertex in the comorbidity graph. To quantify disease co-occurrence in our EHR data selected for constructing the comorbidity graph we use the *Salton's cosine index* (SCI), a cosine similarity measure also known as the Ochiai coefficient (Kalgotra et

al., 2017; Kalgotra, Sharda, & Luse, 2020; Salton & McGill, 1983).

$$\mathrm{SCI}_{uv} = \frac{|A_u \cap A_v|}{\sqrt{|A_u|} \times \sqrt{|A_v|}} \tag{7}$$

We say the co-occurrence of diseases u and v is significant if SCI_{uv} exceeds the user-specified threshold $\Delta \in [0, 1]$ and add the edge $\{u, v\}$ to the comorbidity graph. We use the cosine index to measure similarity (co-occurrence), rather than correlation coefficients that depend explicitly on the sample size (i.e., the number of patients). As the SCI depends only on the prevalence of the diseases in the patient population considered, it is typically more stable and not affected broadly by an increase in the sample size, unless the prevalence of diseases changes significantly.

The comorbidity graph representation we have used is not based on physiological interaction between diseases, which could be more complex and indirect. An edge representing comorbidity instead implies that the two diseases involved co-occur at a relatively high frequency among the patients in the data set. Even if two diseases have indirect interaction physiologically, manifest at different points in time in the progression of a patient, and vary from one patient to another in the data set, they will still correspond to an edge in the comorbidity graph if a sufficiently large number of patients are afflicted with both diseases over the long time window.

In order to choose an appropriate value for the SCI threshold Δ we use a methodology developed in a recent study (Kalgotra et al., 2020). As we cannot directly compute the statistical significance of the cosine index, Kalgotra et al. (2020) use the ϕ-correlation coefficient as a proxy for determining statistical significance. This approach enables us to identify the number of edges with a significance level (α) of 0.01. Given the large sample size, we conservatively used a smaller α to ensure robustness. Our analysis identified 11,086 statistically significant edges. Notably, at an SCI threshold of $\Delta = 0.0484$ the number of significant edges remains the same. Therefore, we choose $\Delta = 0.05$ as our cutoff for including edges in the network. The resulting graph has 285 vertices and 11,086 edges, all contained within one large connected component with 277 vertices.

6. Computational study

The overall goal of our computational study is to assess the performance of the proposed methods and their effectiveness in practice. The experiments we conducted to this end and the results are discussed in this section. All computational experiments reported in this article were conducted on 64-bit Linux® compute nodes with dual Intel® "Skylake" 6130 CPUs with 96 GB RAM. The MILP formulation and the modified BK algorithm were implemented using the Python programming language. Data preparation steps were also implemented in Python. Our source codes are available online (Vaghfi Mohebbi et al., 2024).

First, we compare the performance of the two approaches we have introduced to solve the main optimization problem of interest. In Section 6.1, we report our results from solving MILP formulation (6) using Gurobi Optimization Solver v10.0.3 (Gurobi Optimization, LLC, 2023) on a test bed of relatively smaller instances derived from EHRs. We compare these results against solving the same instances using Algorithm 1, assuming that $C^0 = \varnothing$. In Section 6.2, we focus on the performance of the modified BK algorithm on datasets containing more than 10 million patients, analyze its running

times, and discuss the solutions found. We report our sensitivity analysis experiments with respect to b and solving the marginal mortality rate problem using Algorithm 1 in Section 6.3.

Recall that the denominator of the mortality rate expression must be at least size ℓ, and we use a comparatively relaxed requirement of $\ell = 10$ in the experiments in Section 6.1 and increase it to $\ell = 100$, a stricter requirement for the remaining experiments. Our choices of ℓ have been guided more by our purpose, which is to assess the algorithm performance—smaller values of ℓ imply more feasible cliques and a larger search space. In practice, it is more appropriate to choose ℓ as a sufficiently large fraction of $|P|$ to bestow desired statistical significance on the mortality rate calculation.

Our focus is to discover small ($b \leqslant 4$) disease clusters that exhibit high mortality rates. Based on our conversations with medical practitioners, we understand that these are the cases where it is most valuable, from a decision-support perspective, to find comorbid diseases that increase the mortality rate significantly. In our experiments, we consider the clique size upper-bound $b \in \{1, 2, 3, 4\}$. We note that the largest clique in our comorbidity graph contains 92 diseases, but unsurprisingly there are zero patients that have all 92 diseases. Based on our preliminary experiments, we can detect cliques with as many as 10 vertices in the 10.6 million patient dataset that simultaneously affect at least 100 patients with mortality rate under 0.46.

6.1. Comparisons against the DBC approach

From the 10.6 million patient dataset, for each instance size as represented by the number of patients (1,000, 5,000, 50,000, and 100,000) we selected five random samples that we considered in our experiments comparing the performance of solving the MILP formulation using decomposition and delayed constraint generation against the modified BK algorithm. The delayed addition of constraints (6b)–(6d) described in Section 4.1 was implemented in Gurobi using the "lazy cut" functionality. We set a 3-hour time limit for each instance solved using the DBC algorithm.

At the root node of the branch-and-cut (BC) tree, we begin with a relaxation that does not include any of the constraints (6b)–(6d). Whenever an integral solution to the LP relaxation is encountered at a BC node, all violated constraints (6b)–(6d) are added to the node as lazy cuts. Gurobi re-solves the LP relaxation and manages the subsequent branching at that node.

The detailed results of solving formulation (6) using the DBC algorithm are documented in Table A1 in the appendix and summarized in Table 1. The columns labeled "Size", "b", and "#Optimal" in Table 1 identify the patient size in the test instance used, the clique size upper-bound, and the number of instances solved to optimality, respectively. The next three columns report the average and the range (minimum and maximum) observed over the five random samples of the following quantities (in order): the mortality rate objective function value of the best solution found, the termination MIP gap for instances not solved to optimality, and the wall-clock running time for model building and solution for instances solved to optimality. It is apparent from the results that despite using decomposition and delayed constraint generation, the MILP based approach is unable to solve many of the instances with patient size 50,000 and 100,000 to optimality. By contrast, the modified BK Algorithm 1 is significantly faster on the same test bed and solves all instances to optimality as reported in Table A2 in the appendix and summarized in Table 2. The average wall-clock running

Table 1. Summary of results from using the DBC algorithm to solve formulation (6). Unknown (UNK) means that no value was reported by the solver as the termination was due to a memroy related crash. TO stands for time-out, the solver terminated reaching the 3-hour time limit.

Size	b	#Optimal	Avg μ	[min, max]	Avg gap (%)	[min, max]	Avg time (s)	[min, max]
1,000	1	5	0.30	[0.14, 0.53]	0	-	3	[2, 3]
	2	5	0.45	[0.20, 0.63]	0	-	43	[15, 135]
	3	5	0.48	[0.25, 0.70]	0	-	27	[13, 72]
	4	5	0.51	[0.30, 0.70]	0	-	22	[11, 45]
5,000	1	5	0.63	[0.50, 0.73]	0	-	63	[42, 84]
	2	5	0.68	[0.50, 0.75]	0	-	4,285	[483, 8,895]
	3	4	0.72	[0.60, 0.80]	66	[66, 66]	6,258	[3,457, TO]
	4	3	0.71	[0.60, 0.83]	62	[57, 66]	3,806	[554, TO]
50,000	1	0	0.68	[0.64, 0.72]	45	[36, 56]	TO	TO
	2	0	0.77	[0.67, 0.84]	29	[18, 47]	TO	TO
	3	2	0.91	[0.80, 1.00]	16	[7, 25]	4,022	[2,810, TO]
	4	5	1.00	[1.00, 1.00]	0	-	4,024	[1,060, 9,735]
100,000	1	0	0.46	[0.05, 0.69]	632	[44, 1,800]	TO	TO
	2	0	0.31	[0.03, 0.43]	979	[107, 2,700]	TO	TO
	3	2	0.67	[0.02, 1.00]	3,698	[3,698, 3,698]	TO	[5,032, TO]
	4	2	1.00	[1.00, 1.00]	UNK	[0, UNK]	TO	[4,492, TO]

time reported here includes the time to complete all the recursive calls and the time to report the top-100 highest mortality rate cliques.

From the foregoing experiments it is clear that due to the large number of patients in the data set and the small size of cliques we seek (from a graph with a few hundred vertices), the enumerative approach is dominant in its performance from a computational perspective. Furthermore, this approach also finds the top-K most lethal cliques for our choice of K. Consequently, for our subsequent experiments and studies, we concentrate solely on the modified BK algorithm.

6.2. Scalability of the modified Bron–Kerbosch algorithm

Our primary goal in this section is to assess the performance of the modified BK algorithm on the full dataset consisting of 10.6 million distinct patient records, as it represents the scale at which we may be interested in solving our problems of interest. Table 3 presents the results for this complete dataset. Column "#Dec" reports the number of deceased patients (the numerator of the mortality rate) for reference and column "#RC" reports the number of recursive calls made by the algorithm. For experiments in this section, we report the top-5 most lethal cliques.

As evident from Table 3, the modified BK algorithm is effective on this large dataset for all practically meaningful values of parameter b. As far as the results themselves are concerned, the mortality rate associated with cardiac arrest, either by itself or in combination with other diseases stands out as the highest among all cliques of size at most 4. As expected the maximum mortality rate increases as the upper bound b on clique size increases from 1 to 4. Cardiac arrest by itself has a mortality rate of 0.66, but when paired with coma, septicemia, shock, chronic ulcer, and respiratory distress syndrome, it corresponds to the five highest mortality rate cliques with μ ranging from 0.68 to 0.73. The highest mortality rates for $b \in \{3, 4\}$ are higher, between 0.73 and 0.77, associated with cardiac arrest, coma, and shock.

From the results in Table 3, specifically, the diseases featured in the five highest mortality rate cliques for each value of parameter b, confirm the current understanding

Table 2. Summary of results from solving problem (2) using the modified Bron–Kerbosch algorithm 1.

Size	b	Avg μ	[min, max]	Avg time (s)	[min, max]
1,000	1	0.30	$[0.14, 0.53]$	1	$[1, 1]$
	2	0.45	$[0.20, 0.63]$	1	$[1, 1]$
	3	0.49	$[0.25, 0.70]$	1	$[1, 1]$
	4	0.51	$[0.30, 0.70]$	1	$[1, 1]$
5,000	1	0.63	$[0.50, 0.73]$	1	$[1, 1]$
	2	0.68	$[0.50, 0.75]$	1	$[1, 1]$
	3	0.72	$[0.63, 0.80]$	1	$[1, 1]$
	4	0.74	$[0.63, 0.83]$	5	$[4, 5]$
50,000	1	0.67	$[0.62, 0.72]$	1	$[1, 1]$
	2	0.81	$[0.72, 0.84]$	1	$[1, 1]$
	3	0.98	$[0.91, 1.00]$	7	$[6, 7]$
	4	1.00	$[1.00, 1.00]$	53	$[1, 69]$
100,000	1	0.66	$[0.63, 0.69]$	1	$[1, 1]$
	2	0.79	$[0.74, 0.83]$	1	$[1, 1]$
	3	0.97	$[0.85, 1.00]$	13	$[12, 14]$
	4	1.00	$[1.00, 1.00]$	117	$[114, 121]$

Table 3. Top-5 most lethal cliques in the 10.6 million patient dataset found by the modified BK algorithm.

b	Clique	#Dec	μ	Time (s)	#RC
1	Cardiac arrest	29,450	0.66	20	286
	Shock	23,422	0.41		
	Asp. pneum.	9,677	0.26		
	Resp. fail.	49,599	0.25		
	Malig. neopl.	7,583	0.24		
2	Cardiac arrest, Coma	6,765	0.73	209	11,619
	Cardiac arrest, Septicemia	7,172	0.70		
	Cardiac arrest, Shock	7,176	0.70		
	Cardiac arrest, Chron. ulcer	2,925	0.68		
	Cardiac arrest, Resp. distr. synd.	14,446	0.68		
3	Cardiac arrest, Coma, Shock	90	0.76	1,894	346,017
	Cardiac arrest, Coma, Renal failure	156	0.74		
	Cardiac arrest, Coma, Aftercare	3,991	0.74		
	Cardiac arrest, Coma, Per. atheros.	993	0.74		
	Cardiac arrest, Shock, Septicemia	4,388	0.73		
4	Cardiac arrest, Coma, Shock, Renal failure	1,994	0.77	13,221	7,721,551
	Cardiac arrest, Coma, Shock, Aftercare	1,894	0.77		
	Cardiac arrest, Coma, Shock, Epilepsy	798	0.76		
	Cardiac arrest, Coma, Shock, Diab. mell. w complications	863	0.76		
	Cardiac arrest, Coma, Shock, Fluid disord.	2,469	0.76		

in the medical community as some of the most lethal comorbidities. See for instance, mortality incidence and analysis of these comorbidities by Bauer et al. (2020); Chao et al. (2019); Grubb, Fox, and Elton (1995); Gupte, Knack, and Cramer (2022); Karam et al. (2019); Kempker et al. (2020); Niederman and Cilloniz (2022); Parcha et al. (2021); Vincent, Jones, David, Olariu, and Cadwell (2019); Yan et al. (2020); Yang et al. (2020). Hence, our computational results help reinforce clinical insights through an analysis of large-scale EHR data. Furthermore, the fact that the proposed approach has not detected spurious comorbidities among the highest mortality rate cliques is

encouraging, as the detection of spurious clusters is a common concern in any graph-based data mining approach. The procedure followed in preparing the dataset and the design of the modified BK algorithm seem to be offering a practically effective and useful framework for mortality rate analysis of comorbidities.

Table 4. Top-5 most lethal cliques in the 10.2 million patient dataset found by the modified BK algorithm.

b	Clique	#Dec	μ	Time (s)	#RC
1	Secondary malignancies	24,211	0.23	16	281
	Cancer of liver	3,189	0.20		
	Cancer of pancreas	2,376	0.20		
	Intrauterine hypoxia	125	0.20		
	Cancer of lung	11,565	0.19		
2	Septicemia, Secondary malignancies	7,380	0.51	268	11,203
	Septicemia, Cancer of lung	3,144	0.47		
	Coma, Myocard. infarc.	2,695	0.47		
	Secondary malignancies, Pneumonia	7,866	0.46		
	Secondary malignancies, Renal failure	7,734	0.45		
3	Cancer of lung, Septicemia, Secondary malignancies	2,249	0.58	2,011	329,843
	Coag. disord., Secondary malignancies, Septicemia	2,825	0.57		
	Secondary malignancies, Renal failure, Septicemia	3,864	0.57		
	Pneumonia, Secondary malignancies, Septicemia	3,663	0.57		
	Nutritional deficiencies, Secondary malignancies, Septicemia	3,666	0.56		
4	Septicemia, Renal failure, Cancer of lung, Second. malign.	991	0.62	11,901	7,291,810
	Septicemia, Coag. disord., Gastrointestinal hemor., Second. malign.	745	0.62		
	Septicemia, Coag. disord., Cancer of lung, Second. malign.	796	0.62		
	Septicemia, Cancer of lung, Pulmonary heart dis., Second. malign.	483	0.61		
	Septicemia, Coag. disord., Pneumonia, Second. malign.	1,461	0.61		

In order to explore the dataset further, we exclude from the 10.6 million patient dataset, every patient that has cardiac arrest, shock, aspiration pneumonitis, respiratory failure, or malignant neoplasm, which may be evident to physicians as causing lethal comorbidities. The resulting dataset corresponding to 10.2 million patients is analyzed using the modified BK algorithm and the results are reported in Table 4.

From the results in Table 4, the individual mortality rate linked to secondary malignancies is the highest among all the (remaining) diseases, although it is much lower at 0.23. Notably, when paired with septicemia it corresponds to the highest mortality rate of 0.51, more than doubling mortality rate of secondary malignancies and more than the sum of the individual mortality rates. These two diseases continue to remain among the top-two highest mortality rate cliques when we increase b to three and then to four, with the respective rates increasing to 0.58 and then to 0.62. This perspective into lethal comorbidities derived from the EHR dataset after excluding the diseases that dominate the most lethal cliques illustrates how the proposed framework may be used to better understand lesser known comorbidities. This is especially important given how the maximum mortality rate rapidly escalates in the presence of comorbidities, even though it may not occur intuitively to practitioners.

The findings presented in Tables 3 and 4 demonstrate that the modified BK algorithm tackled these large-scale datasets within a reasonably short span of less than three hours in computation time. These results offer compelling evidence that our proposed methodology seamlessly integrates with real-world datasets, underscoring its robustness and effectiveness in practical healthcare analytics settings.

6.3. Sensitivity analysis and marginal mortality rates

Our experiments in Sections 6.1 and 6.2 focused on solving problem (2), while the experiments in this section focus on problem (3) when C^0 is not empty. For experiments in this section, we consider the 10.2 million patient dataset described in the previous section that excludes every patient with any of the following diseases: cardiac arrest, shock, aspiration pneumonitis, respiratory failure, and malignant neoplasm. From Table 4, for each value of b, we fix the fifth highest mortality rate clique as C^0. We choose C^0 in this manner as the results in Tables 3 and 4 already demonstrate a sequence of containment relationships among the cliques as b is increased. With this choice of C^0, we aim to identify high marginal mortality rate cliques not readily apparent from the results in Table 4.

We first report on our experiments on the impact of adding a new disease $u \in V \backslash C^0$ on the mortality rate. In this experiment, we do not require that $C^0 \cup \{u\}$ forms a clique. Then, we evaluate the modified BK algorithm when it is tasked to find maximum marginal mortality rate cliques. Specifically, we conduct the following experiments with each C^0 chosen as described above: (i) For each $u \in V \backslash C^0$, compute the mortality rate of $C^0 \cup \{u\}$ and return the top-3 highest marginal mortality rate *subsets* containing C^0. (ii) Find the top-3 highest marginal mortality rate *cliques* with respect to to C^0 using Algorithm 1 when allowing the addition of one and two more diseases.

Table 5. Top-3 most lethal subsets formed by the addition of a single disease.

Initial clique C^0	$\mu(C^0)$	Additional disease u	$\mu(C^0 \cup \{u\})$	%Increase	Cluster type
Cancer of lung	0.19	Coma	0.51	168%	1-defective clique
	0.19	Septicemia	0.47	147%	clique
	0.19	Chronic ulcer of skin	0.40	110%	1-defective clique
Secondary malignancies Renal failure	0.45	Coma	0.63	40%	1-defective clique
	0.45	Septicemia	0.57	26%	clique
	0.45	Cancer of esophagus	0.55	22%	2-defective clique
Septicemia Secondary malignancies Nutritional deficiencies	0.56	Cancer of esophagus	0.66	17%	2-defective clique
	0.56	Coma	0.65	16%	1-defective clique
	0.56	Cancer of lung	0.60	7%	clique
Septicemia Secondary malignancies Pneumonia Coag. disorder.	0.61	Coma	0.67	9%	1-defective clique
	0.61	Cancer of pancreas	0.65	6%	3-defective clique
	0.61	Chronic ulcer of skin	0.65	6%	1-defective clique

Table 5 reports the results from our one-sided sensitivity analysis with respect to the clique size upper-bound being increased by one. For each C^0 considered in the first column, we report the additional disease u that corresponds to the highest, second highest and the third highest increase in mortality rate in the third column. Column labeled "%Increase" reports the percentage by which $\mu(C^0 \cup \{u\})$ has increased over $\mu(C^0)$. In this test instance, we observe that when C^0 is smaller, $\mu(C^0)$ is typically smaller and significant increases are possible (and observed) by the addition of a single disease. For example, adding coma to the initial clique containing just one disease, cancer of lung, increases mortality rate by 168%. Conversely, when C^0 is larger, $\mu(C^0)$ is already quite high and the observed increases from the addition of another disease are not as significant. For example, when $|C^0| = 4$, the largest increase in

mortality rate is less than 10%. These observations also reiterate our emphasis on small but lethal cliques in this article and substantiate our parameter choice in conducting computational experiments with $b \leqslant 4$.

Another key observation concerns the types of subsets we detect corresponding to the three highest marginal mortality rates, as indicated under the column labeled "Cluster type". We characterize them using a graph-theoretic clique relaxation known as a k-defective clique, which is a subset of vertices that induces a subgraph that falls short of containing all possible edges by at most k edges. Formally, $S \subset V$ is a k-defective clique if the induced subgraph $G[S]$ contains at least $\binom{|S|}{2} - k$ edges. A classical clique is therefore 0-defective. We characterize the subsets we find in terms of defective cliques as they are, in our opinion, a more intuitive choice in this context. Other authors, for example, may choose to characterize these using other clique relaxations like k-plexes (Balasundaram, Butenko, & Hicks, 2011) or γ-quasi-cliques (Pattillo, Veremyev, Butenko, & Boginski, 2013) with an appropriate choice of the parameters k and γ.

Our analysis reveals that 9 out of the 12 subsets are 0-defective or 1-defective cliques, although when a 2-vertex subset is 1-defective, it is simply two isolated vertices, which is arguably not a very interesting cluster. However, 9 out of the 12 subsets induce connected subgraphs and we visualize three such examples in Figure 2. It is worth reiterating that an edge in the comorbidity graph is constructed based on co-occurrence between the endpoints in the EHR dataset under consideration. Hence, for two diseases $u, v \in V$, it is possible that $\mathrm{SCI}_{uv} < \Delta$ while $|A_u \cap A_v| \geqslant \ell$, permitting a subset of vertices to contain non-adjacent diseases although it corresponds to a mortality rate strictly larger than zero. We remark that the choice of $\ell = 100$ in our experiments is arbitrary, and larger values of lower-bound ℓ may yield different clusters.

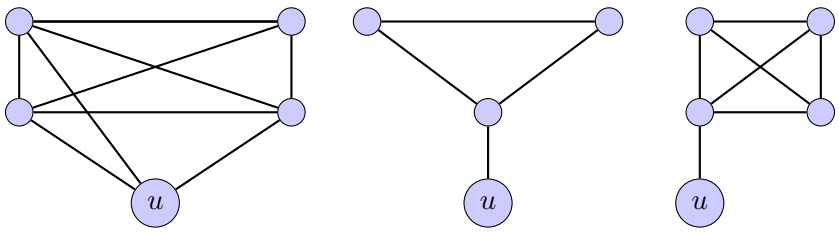

Figure 2. From left to right, we visualize the subgraphs induced by the following subsets identified in Table 5: 1-defective clique {Septicemia, Secondary malignancies, Pneumonia, Coag disorder, Coma (u)}; 2-defective clique {Septicemia, Secondary malignancies, Nutritional deficiencies, Cancer of esophagus (u)}; 3-defective clique {Septicemia, Secondary malignancies, Pneumonia, Coag disorder, Cancer of pancreas (u)}. The added vertex u is labeled in the graphs.

Our final experiments assess the performance of Algorithm 1 given a non-empty C^0. The modified BK algorithm is effective on the large-scale dataset we have considered, and it is extremely fast when an initial clique of diseases C^0 is provided. We can also see from the results in Tables 6 and 7 that this experiment offers information complementary to that reported in Tables 4, and we observe trends similar to Table 5. Through this experiment we can recognize the disease(s) to be most wary of when a patient is already diagnosed with an existing clique of diseases, which is the main purpose behind investigating marginal mortality rates.

Table 6. Top-3 most lethal cliques formed by the addition of a single disease.

C^0	$\mu(C^0)$	Additional disease u	$\mu(C^0 \cup \{u\})$	%Increase	Time (s)	#RC
Cancer of lung	0.19	Septicemia	0.47	147%	2	48
	0.19	Renal failure	0.40	110%		
	0.19	Pneumonia	0.38	100%		
Secondary malignancies Renal failure	0.45	Septicemia	0.57	26%	2	75
	0.45	Pneumonia	0.54	20%		
	0.45	Cancer of liver	0.53	17%		
Septicemia Secondary malignancies Nutritional deficiencies	0.56	Cancer of lung	0.60	11%	2	70
	0.56	Coag. disord.	0.60	11%		
	0.56	Pneumonia	0.60	11%		
Septicemia Secondary malignancies Pneumonia Coag. disorder.	0.61	Gastrointestinal hemor.	0.64	4%	2	69
	0.61	Intest. obstruc. w hernia	0.64	4%		
	0.61	Renal failure	0.63	3%		

Table 7. Top-3 most lethal cliques formed by the addition of two diseases.

C^0	$\mu(C^0)$	Additional diseases u, v	$\mu(C^0 \cup \{u, v\})$	%Increase	Time (s)	#RC
Cancer of lung	0.19	Secondary malignancies, Septicemia	0.57	200%	3	959
	0.19	Coag. disord., Septicemia	0.55	189%		
	0.19	Nutritional deficiencies, Septicemia	0.53	178%		
Secondary malignancies Renal failure	0.45	Cancer of lung, Septicemia	0.62	37%	3	2454
	0.45	Coag. disord., Septicemia	0.61	35%		
	0.45	Pneumonia, Septicemia	0.60	33%		
Septicemia Secondary malignancies Nutritional deficiencies	0.56	Gastrointestinal hemorr., Coag. disord.	0.64	14%	3	2318
	0.56	Renal failure, Cancer of lung	0.63	12%		
	0.56	Liver disease, Renal failure	0.63	12%		
Septicemia Secondary malignancies Pneumonia Coag. disorder.	0.61	Gastrointestinal hemor., Bone disease	0.69	13%	2	2220
	0.61	Gastrointestinal hemor., Phlebitis	0.68	11%		
	0.61	Intest. obstruc. w hernia, E Code: medical drugs	0.67	9%		

7. Conclusion

This article introduces a framework for an integrated analysis of an EHR dataset in conjunction with a comorbidity graph to detect lethal cliques, i.e., co-occurring disease clusters with high mortality rates. Our computational study demonstrates the effectiveness of the proposed methodology on large-scale EHR datasets using the enumerative algorithm we introduce to discover the most lethal cliques containing up to four vertices in the comorbidity graph. Our results, specifically the most lethal comorbidities identified, are consistent with those observed in the medical literature.

Identifying such disease clusters with high (marginal) mortality rates can empower healthcare providers to implement early interventions and allocate special attention to patients afflicted by these diseases, thereby enhancing healthcare outcomes. When physicians are aware of a patient's preexisting comorbidities, it becomes valuable to determine which additional diseases may further increase mortality rates. This insight can enable patients and healthcare practitioners to take proactive measures to mitigate the risk.

By filtering the EHR dataset appropriately, we can study, for example, specific age groups or demographics to better understand their exposure to lethal comorbidi-

ties. Furthermore, as we demonstrate in Table 4, our framework enables us to study lesser known comorbidities by excluding from the input data, diseases that dominate the highest mortality rate clusters, and consequently are already well understood. Combined with the domain expertise of clinicians' specialties, the access to massive datasets being generated through electronic health records presents an opportunity for healthcare analytics researchers to work with clinicians and medical researchers to validate lethal comorbidities observed in practice, and perhaps discover previously unknown comorbidities and insights into higher mortality rates of specific patient subgroups.

Limitations of the proposed methodology and results. Our dataset is limited to patients from hospitals using the Cerner EHR system, which may introduce biases related to the healthcare delivery practices and patient populations of these institutions. Additionally, data may be incomplete if patients did not adhere to follow-up appointments, sought care at facilities utilizing different EHR systems, or received treatment outside of the Cerner network. Furthermore, during data preprocessing, we excluded patients with missing information required for a mortality rate analysis, such as diagnosis codes, time-stamps of encounters, and discharge status. This might affect the representativeness of our sample, although the large size of the final dataset mitigates this risk to a large extent. Although our dataset contains a significantly large number of patients with their most recent five-year medical history and our findings have been consistent with existing medical literature, external validation using diverse datasets across different EHR systems and healthcare settings would support the generalizability of our methodology and enhance the robustness of our results.

Future extensions. It may be worth extending our methodology to alternative cluster models that are better suited if the EHR database used to construct the comorbidity graph is incomplete, possibly because of short time windows or if an excessive number of encounters are suspected to be missing. Clique relaxations like k-plexes, robust 2-clubs, quasi-cliques, and defective cliques allow non-adjacent pairs of vertices inside a cluster and are less sensitive to missing edges; for an introduction to such models, see (Balasundaram et al., 2011; Y. Lu, Salemi, Balasundaram, & Buchanan, 2022; Miao & Balasundaram, 2020; Pattillo, Youssef, & Butenko, 2013; Trukhanov, Balasubramaniam, Balasundaram, & Butenko, 2013). The scope of this work can be further expanded by an explicit accounting of the temporal nature of the patient encounters using temporal comorbidity graphs introduced by Y. Lu et al. (2021).

Consent and approval

This work was conducted with the electronic health records database from the Cerner Health Facts®, de-identified and made available to the Center for Health Systems Innovation at Oklahoma State University. Any opinions, findings, conclusions, or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the Oracle Cerner Corporation. The Institutional Review Board at Oklahoma State University Center for Health Sciences exempted the study from review.

Disclosure statement

The authors report there are no competing interests to declare.

Funding

No specific funding was received for this work.

Acknowledgement

The authors thank the High Performance Computing Center at Oklahoma State University, supported in part through the National Science Foundation grant OAC-1531128, for providing the necessary computing support. The authors also thank the reviewers for the feedback that helped us improve our presentation.

References

Agency for Healthcare Research and Quality (AHRQ). (2021). *Software. Healthcare Cost and Utilization Project (HCUP).* Rockville, MD: Agency for Healthcare Research and Quality.

Ahluwalia, S. C., Gross, C. P., Chaudhry, S. I., Ning, Y. M., Leo-Summers, L., Van Ness, P. H., & Fried, T. R. (2012). Impact of comorbidity on mortality among older persons with advanced heart failure. *Journal of General Internal Medicine*, *5*(267), 513-519.

Arias, E., & Xu, J. (2022). *United States life tables, 2020* (National Vital Statistics Reports No. vol 71 no 1). Hyattsville, MD: National Center for Health Statistics (U.S.). (`https://stacks.cdc.gov/view/cdc/118055`)

Balasundaram, B., Butenko, S., & Hicks, I. V. (2011). Clique relaxations in social network analysis: The maximum k-plex problem. *Operations Research*, *59*(1), 133–142.

Baud, D., Qi, X., Nielsen-Saines, K., Musso, D., Pomar, L., & Favre, G. (2020). Real estimates of mortality following covid-19 infection. *The Lancet Infectious Diseases*, *20*(7), 773.

Bauer, M., Gerlach, H., Vogelmann, T., Preissing, F., Stiefel, J., & Adam, D. (2020). Mortality in sepsis and septic shock in Europe, North America and Australia between 2009 and 2019—results from a systematic review and meta-analysis. *Critical Care*, *24*(1), 239.

Braunstein, J. B., Anderson, G. F., Gerstenblith, G., Weller, W., Niefeld, M., Herbert, R., & Wu, A. W. (2003). Noncardiac comorbidity increases preventable hospitalizations and mortality among medicare beneficiaries with chronic heart failure. *Journal of the American College of Cardiology*, *7*(42), 1226-1233.

Bron, C., & Kerbosch, J. (1973). Algorithm 457: Finding all cliques on an undirected graph. *Communications of ACM*, *16*, 575-577.

Cazals, F., & Karande, C. (2008). A note on the problem of reporting maximal cliques. *Theoretical Computer Science*, *407*(1), 564–568.

Centers for Disease Control and Prevention. (2013). *International classification of diseases, 9th revision, clinical modification (ICD-9-CM).* Retrieved from `https://archive.cdc.gov/#/details?url=https://www.cdc.gov/nchs/icd/icd9cm.htm`

Centers for Disease Control and Prevention. (2022). *International classification of diseases, 10th revision, clinical modification (ICD-10-CM).* Retrieved from `https://www.cdc.gov/nchs/icd/icd-10-cm/`

Chao, C., Bhatia, S., Xu, L., Cannavale, K. L., Wong, F. L., Huang, P.-Y. S., ... Armenian, S. H. (2019). Incidence, risk factors, and mortality associated with second malignant neoplasms among survivors of adolescent and young adult cancer. *JAMA Network Open*, *2*(6), e195536-e195536.

Chen, J., Normand, S.-L. T., Wang, Y., & Krumholz, H. M. (2011). National and regional trends in heart failure hospitalization and mortality rates for medicare beneficiaries, 1998-2008. *JAMA*, *306*(15), 1669–1678.

Chiba, N., & Nishizeki, T. (1985). Arboricity and subgraph listing algorithms. *SIAM Journal on Computing*, *14*(1), 210-223.

Corraini, P., Szépligeti, S., Henderson, V., Ording, A., Horváth-Puhó', E., & rensen, H. S. (2018). Comorbidity and the increased mortality after hospitalization for stroke: A population-based cohort study. *Journal of Thrombosis and Haemostasis*, *16*(2), 242-252.

Eppstein, D., Löffler, M., & Strash, D. (2013). Listing all maximal cliques in large sparse real-world graphs. *ACM Journal of Experimental Algorithmics*, *18*, 3.1–3.21.

Feinstein, A. R. (1970). The pre-therapeutic classification of co-morbidity in chronic disease. *Journal of Chronic Diseases*, *23*(7), 455-468.

Gijsen, R., Hoeymans, N., Schellevis, F. G., Ruwaard, D., Satariano, W. A., & van den Bos, G. A. (2001). Causes and consequences of comorbidity: A review. *Journal of Clinical Epidemiology*, *54*(7), 661-674.

Grubb, N., Fox, K., & Elton, R. (1995). In-hospital mortality after out-of-hospital cardiac arrest. *The Lancet*, *346*(8972), 417-421.

Gupte, T., Knack, A., & Cramer, J. D. (2022). Mortality from aspiration pneumonia: Incidence, trends, and risk factors. *Dysphagia*, *37*(6), 1493–1500.

Gurobi Optimization, LLC. (2023). *Gurobi Optimizer Reference Manual.* https://www.gurobi.com.

Hidalgo, C. A., Blumm, N., Barabási, A.-L., & Christakis, N. A. (2009). A dynamic network approach for the study of human phenotypes. *PLOS Computational Biology*, *5*(4), 1–11.

Holguin, F., Folch, E., Redd, S. C., & Mannino, D. M. (2005). Comorbidity and mortality in COPD-related hospitalizations in the United States, 1979 to 2001. *Chest*, *4*(128), 2005-2011.

Kalgotra, P., Sharda, R., & Croff, J. M. (2017). Examining health disparities by gender: A multimorbidity network analysis of electronic medical record. *International Journal of Medical Informatics*, *108*, 22–28.

Kalgotra, P., Sharda, R., & Luse, A. (2020). Which similarity measure to use in network analysis: Impact of sample size on phi correlation coefficient and Ochiai index. *International Journal of Information Management*, *55*, 102229.

Kansal, A., Gao, M., Balu, S., Nichols, M., Corey, K., Kashyap, S., & Sendak, M. (2021). Impact of diagnosis code grouping method on clinical prediction model performance: a multi-site retrospective observational study. *International Journal of Medical Informatics*, *151*, 104466.

Karam, N., Bataille, S., Marijon, E., Tafflet, M., Benamer, H., Caussin, C., ... Lambert, Y. (2019). Incidence, mortality, and outcome-predictors of sudden cardiac arrest complicating myocardial infarction prior to hospital admission. *Circulation: Cardiovascular Interventions*, *12*(1), e007081.

Kempker, J. A., Abril, M. K., Chen, Y., Kramer, M. R., Waller, L. A., & Martin, G. S. (2020). The epidemiology of respiratory failure in the united states 2002–2017: A serial cross-sectional study. *Critical Care Explorations*, *2*(6), e0128.

Koskinen, M., Salmi, J. K., Loukola, A., Mäkelä, M. J., Sinisalo, J., Carpén, O., & Renkonen, R. (2022). Data-driven comorbidity analysis of 100 common disorders reveals patient subgroups with differing mortality risks and laboratory correlates. *Scientific Reports*, *12*(1), 18492.

Lee, S. H., Levin, D., Finley, P. D., & Heilig, C. M. (2019). Chief complaint classification with recurrent neural networks. *Journal of Biomedical Informatics*, *93*, 103158.

Leontiadis, G. I., Molloy-Bland, M., Moayyedi, P., & Howden, C. W. (2013). Effect of comorbidity on mortality in patients with peptic ulcer bleeding: systematic review and meta-analysis. *The American Journal of Gastroenterology*, *108*(3), 331–346.

Lu, W. S., & Tsutakawa, R. K. (1996). Analysis of mortality rates via marginal extended quasi-likelihood. *Statistics in Medicine*, *15*(13), 1397–1407.

Lu, Y., Chen, S., Miao, Z., Delen, D., & Gin, A. (2021). Clustering temporal disease networks

to assist clinical decision support systems in visual analytics of comorbidity progression. *Decision Support Systems*, *148*, 113583.

Lu, Y., Salemi, H., Balasundaram, B., & Buchanan, A. (2022). On fault-tolerant low-diameter clusters in graphs. *INFORMS Journal on Computing*, *34*(6), 3181–3199.

Malecki, S. L., Loffler, A., Tamming, D., Johansen, N. D., Biering-Sørensen, T., Fralick, M., ... others (2024). Development and external validation of tools for categorizing diagnosis codes in international hospital data. *International Journal of Medical Informatics*, 105508.

Marrie, R. A., Elliott, L., Marriott, J., Cossoy, M., Blanchard, J., Leung, S., & Yu, N. (2015). Effect of comorbidity on mortality in multiple sclerosis. *Neurology*, *85*(3), 240–247.

Marshall, R. J. (2018). Mapping disease and mortality rates using empirical Bayes estimators. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *40*(2), 283-294.

Miao, Z., & Balasundaram, B. (2020). An ellipsoidal bounding scheme for the quasi-clique number of a graph. *INFORMS Journal on Computing*, *32*(3), 763–778.

Niederman, M. S., & Cilloniz, C. (2022). Aspiration pneumonia. *Revista espanola de quimioterapia*, *35*(Suppl 1), 73-77.

Osmond, C. (1985). Using Age, Period and Cohort Models to Estimate Future Mortality Rates. *International Journal of Epidemiology*, *14*(1), 124–129.

Parcha, V., Kalra, R., Bhatt, S. P., Berra, L., Arora, G., & Arora, P. (2021). Trends and geographic variation in acute respiratory failure and ARDS mortality in the United States. *Chest*, *159*(4), 1460–1472.

Pattillo, J., Veremyev, A., Butenko, S., & Boginski, V. (2013). On the maximum quasi-clique problem. *Discrete Applied Mathematics*, *161*(1–2), 244–257. (10.1016/j.dam.2012.07.019)

Pattillo, J., Youssef, N., & Butenko, S. (2013). On clique relaxation models in network analysis. *European Journal of Operational Research*, *226*(1), 9–18.

Porta, M. (Ed.). (2014). *A dictionary of epidemiology* (5th ed.). Oxford: Oxford University Press.

Redelmeier, D. A., Tan, S. H., & Booth, G. L. (1998). The treatment of unrelated disorders in patients with chronic medical diseases. *The New England Journal of Medicine*, *21*(338), 1516–1520.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York, NY, USA: McGraw-Hill, Inc.

San Segundo, P., Artieda, J., & Strash, D. (2018). Efficiently enumerating all maximal cliques with bit-parallelism. *Computers & Operations Research*, *92*, 37–46.

Short, S. E., Yang, Y. C., & Jenkins, T. M. (2013). Sex, gender, genetics, and health. *American Journal of Public Health*, *103*, S93-S101.

Tomita, E., Tanaka, A., & Takahashi, H. (2006). The worst-case time complexity for generating all maximal cliques and computational experiments. *Theoretical Computer Science*, *363*(1), 28–42.

Trukhanov, S., Balasubramaniam, C., Balasundaram, B., & Butenko, S. (2013). Algorithms for detecting optimal hereditary structures in graphs, with application to clique relaxations. *Computational Optimization and Applications*, *56*(1), 113-130.

Vaghfi Mohebbi, P., Lu, Y., Miao, Z., Balasundaram, B., Kalgotra, P., & Sharda, R. (2024). *Implementation of a modified Bron–Kerbosch algorithm and an MILP formulation for the maximum mortality rate clique problem.* Codes online at: `https://github.com/ParisaMohebbi/MaxMortClique`.

Van Gestel, Y. R. B. M., Lemmens, V. E. P. P., de Hingh, I. H. J. T., Steevens, J., Rutten, H. J. T., Nieuwenhuijzen, G. A. P., ... Siersema, P. D. (2013). Influence of comorbidity and age on 1-, 2-, and 3-month postoperative mortality rates in gastrointestinal cancer patients. *Annals of Surgical Oncology*, *20*(2), 371–380.

Vincent, J.-L., Jones, G., David, S., Olariu, E., & Cadwell, K. K. (2019). Frequency and mortality of septic shock in Europe and North America: A systematic review and meta-analysis. *Critical Care*, *23*(1), 196.

Xu, J., Murphy, S. L., Kochanek, K. D., & Arias, E. (2016, December). *Mortality in the United States, 2015* (NCHS Data Brief No. 267). Hyattsville, MD: National Center for

Health Statistics.

Yan, S., Gan, Y., Jiang, N., Wang, R., Chen, Y., Luo, Z., ... Lv, C. (2020). The global survival rate among adult out-of-hospital cardiac arrest patients who received cardiopulmonary resuscitation: a systematic review and meta-analysis. *Critical Care*, *24*(1), 61.

Yang, W. S., Kang, H. D., Jung, S. K., Lee, Y. J., Oh, S. H., Kim, Y.-J., ... Kim, W. Y. (2020). A mortality analysis of septic shock, vasoplegic shock and cryptic shock classified by the third international consensus definitions (sepsis-3). *The Clinical Respiratory Journal*, *14*(9), 857-863.

Zolbanin, H. M., Delen, D., & Zadeh, A. H. (2015). Predicting overall survivability in comorbidity of cancers: A data mining approach. *Decision Support Systems*, *74*, 150-161.

Appendix A. Detailed computational results

Table A1 presents the results from solving formulation (6) using the DBC algorithm. The columns labeled "Size", "No.", and "b" identify the patient size in the instance, the sample number, and clique size upper-bound, respectively. Under the column labeled "Clique" we report the diseases corresponding to the best solution found and under column "μ" we report the corresponding mortality rate. The column labeled "Gap (%)" lists the MILP termination gap reported by the solver, and it is zero whenever the solver finds an optimal solution. Under the column labeled "Time (s)" we report the wall-clock running time in seconds, rounded down to the nearest integer, for building the MILP relaxation used at the root node and solving it using the DBC algorithm. The entry "TO" denotes that the solver timed out, exceeding the 3-hour time limit.

The detailed results from using the modified BK Algorithm 1 to solve problem (2) is presented in Table A2, with the columns labeled as before. The wall-clock running time reported here includes the time to complete all the recursive calls and the time to report the top-100 highest mortality rate cliques. Although top-100 highest mortality rate cliques were found, for comparison purposes, we only report the maximum mortality rate clique found in Table A2.

Table A1. Detailed results of using the DBC algorithm to solve formulation (6).

Size	No.	b	Clique	μ	Gap (%)	Time (s)
1K	1	1	Coma	0.14	-	2
		2	Skin tissue infect., Defic. and other anemia	0.20	-	33
		3	Urinary tract infect., Unclass., Disord. lipid metabol	0.25	-	13
		4	Urinary tract infect., Unclass., Disord. lipid metabol, Defic. and other anemia	0.30	-	11
	2	1	Renal fail.	0.20	-	3
		2	Lower resp. dis., Renal fail.	0.40	-	15
		3	Lower resp. dis., Renal fail.	0.40	-	15
		4	Lower resp. dis., Renal fail.	0.40	-	16
	3	1	Resp. fail.	0.32	-	3
		2	Lower resp. dis., Dis. of white blood cells	0.50	-	15
		3	Resp. fail., Nerv. sys. disord., Disord. lipid metabol	0.70	-	16
		4	Resp. fail., Nerv. sys. disord., Disord. lipid metabol	0.70	-	19
	4	1	Resp. fail.	0.53	-	3
		2	Resp. fail., Fluid disord.	0.63	-	135
		3	Resp. fail., Fluid disord.	0.63	-	72
		4	Resp. fail., Fluid disord.	0.63	-	45
	5	1	Resp. fail.	0.31	-	3
		2	Chronic obs. pulmonary dis., Renal fail.	0.50	-	19
		3	Liver diseas., Gastrointestinal disord., Renal fail.	0.50	-	18
		4	Liver diseas., Fever unknown origin	0.50	-	17
5K	1	1	Cardiac arrest	0.60	-	84
		2	Resp. fail., Cardiac arrest	0.75	-	8895
		3	Shock, Resp. fail., Chronic kidney dis.	0.80	-	5932
		4	Septicemia, Resp. fail., Liver diseas., Renal fail.	0.83	-	6725
	2	1	Cardiac arrest	0.70	-	42
		2	Coronary atherosclerosis, Cardiac arrest	0.72	-	2931
		3	Coronary atherosclerosis, Cardiac arrest	0.72	-	5277
		4	Resp. fail., Lower resp. dis., Nutri. defic., Defic. and other anemia	0.60	66	TO
	3	1	Cardiac arrest	0.50	-	69
		2	Resp. fail., Upper resp. infect.	0.50	-	3584
		3	Pneumonia, Circulatory dis., Dis. of white blood cells	0.60	66	TO
		4	Secondary malign., Lower. Resp. dis., Aftercare	0.63	57	TO
	4	1	Cardiac arrest	0.73	-	63
		2	Cardiac arrest	0.73	-	5533
		3	Cardiac arrest	0.73	-	3457
		4	Cardiac arrest	0.73	-	4141
	5	1	Cardiac arrest	0.62	-	58
		2	Shock, Mental health	0.70	-	483
		3	Urinary tract infect., Septicemia, Resp. fail.	0.76	-	10368
		4	Urinary tract infect., Resp. fail., Pneumonia, Abdominal pain	0.80	-	554
50K	1	1	-	-	-	TO
		2	Shock, Coma	0.67	47	TO
		3	Septicemia, Mainten. chemotherapy, Genitourinary symptoms	0.80	25	TO
		4	Shock, Metabolic disord., Coma, Cardiac arrest	1.00	-	3792
	2	1	Cardiac arrest	0.72	36	TO
		2	Dis. of white blood cells, Cardiac arrest	0.84	18	TO
		3	Septicemia, Dis. of white blood cells, Cardiac arrest	1.00	-	5235
		4	Resp. fail., Coma, Chronic obs. pulmonary dis., Cardiac arrest	1.00	-	9735
	3	1	Cardiac arrest	0.64	56	TO
		2	Coma, Cardiac arrest	0.81	22	TO
		3	Shock, Coag. disord., Cardiac arrest	0.93	7	TO
		4	Diab. mell. with complications, Coma, Cardiac arrest, Renal fail.	1.00	-	1060
	4	1	Cardiac arrest	0.70	42	TO
		2	Epilepsy, Cardiac arrest	0.83	20	TO
		3	Nutri. defic., Renal fail., Cardiac arrest	1.00	-	2810
		4	Essential hypertension, Epilepsy, Cardiac arrest, Renal fail.	1.00	-	2185
	5	1	-	-	-	TO
		2	Chronic obs. pulmonary dis., Cardiac arrest	0.72	37	TO
		3	Septicemia, Mainten. chemotherapy, E Codes. effects of medical drugs	0.86	15	TO
		4	Secondary malign., Pneumonia, Septicemia, Cardiac arrest	1.00	-	3349
100K	1	1	Cardiac arrest	0.69	44	TO
		2	Disord. in infancy	0.03	2700	TO
		3	Nutri. defic., Cardiac arrest, Myocardial infarc.	1.00	-	5,032
		4	-	-	-	TO
	2	1	Coag. disord.	0.05	1800	TO
		2	-	-	-	TO
		3	Phlebitis, Chronic ulcer skin, Cardiac arrest	1.00	-	7,954
		4	Shock, Endocrine disord., Congest. heart fail., Biliary tract dis.	1.00	-	4,492
	3	1	Cardiac arrest	0.65	52	TO
		2	Shock, Biliary tract dis.	0.43	130	TO
		3	Unclass.	0.02	3698	TO
		4	Nutri. defic., Disord. lipid metabol, Dis. of white blood cells, Cardiac arrest	1.00	-	5,517
	4	1	-	-	-	TO
		2	-	-	-	TO
		3	-	-	-	TO
		4	-	-	-	TO
	5	1	-	-	-	TO
		2	Malign. neopl. without site, E Codes. medical care	0.48	107	TO
		3	-	-	-	TO
		4	-	-	-	TO

Table A2. Detailed results of using the modified Bron–Kerbosch Algorithm 1 to solve problem (2).

Sample	No.	b	Clique	μ	Time (s)
1K	1	1	Coma	0.14	<1
		2	Skin tissue infect., Defic. and other anemia	0.20	<1
		3	Urinary tract infect., Unclass., Disord. lipid metabol	0.25	<1
		4	Urinary tract infect., Unclass., Disord. lipid metabol, Defic. and other anemia	0.30	<1
	2	1	Renal fail.	0.20	<1
		2	Lower. Resp. dis., Renal fail.	0.40	<1
		3	Urinary tract infect., Metabolic disord., Cardiac dysrhythmias	0.40	<1
		4	Lower. Resp. dis., Renal fail.	0.40	<1
	3	1	Resp. fail.	0.31	<1
		2	Lower. Resp. dis., Dis. of white blood cells	0.50	<1
		3	Resp. fail., Nerv. sys. disord., Disord. lipid metabol	0.70	<1
		4	Resp. fail., Unclass., Nerv. sys. disord., Disord. lipid metabol	0.70	<1
	4	1	Resp. fail.	0.53	<1
		2	Resp. fail., Fluid disord.	0.63	<1
		3	Resp. fail., Fluid disord.	0.63	<1
		4	Resp. fail., Fluid disord.	0.63	<1
	5	1	Resp. fail.	0.31	<1
		2	Chronic obs. pulmonary dis., Renal fail.	0.50	<1
		3	Liver diseas., Gastrointestinal disord., Renal fail.	0.50	<1
		4	Liver diseas., Gastrointestinal disord., Renal fail.	0.50	<1
5K	1	1	Cardiac arrest	0.60	<1
		2	Resp. fail., Cardiac arrest	0.75	<1
		3	Shock, Resp. fail., Chronic kidney dis.	0.80	1
		4	Septicemia, Resp. fail., Liver diseas., Renal fail.	0.83	5
	2	1	Cardiac arrest	0.70	<1
		2	Coronary atherosclerosis, Cardiac arrest	0.72	<1
		3	Coronary atherosclerosis, Cardiac arrest	0.72	<1
		4	Coronary atherosclerosis, Cardiac arrest	0.72	5
	3	1	Cardiac arrest	0.50	<1
		2	Resp. fail., Upper resp. infect.	0.50	<1
		3	Secondary malign., Lower. Resp. dis., Aftercare	0.63	<1
		4	Secondary malign., Lower. Resp. dis., Aftercare	0.63	5
	4	1	Cardiac arrest	0.73	<1
		2	Cardiac arrest	0.73	<1
		3	Cardiac arrest	0.73	<1
		4	Cardiac arrest	0.73	4
	5	1	Cardiac arrest	0.62	<1
		2	Shock, Mental health	0.70	<1
		3	Urinary tract infect., Septicemia, Resp. fail.	0.76	<1
		4	Urinary tract infect., Resp. fail., Pneumonia, Abdominal pain	0.80	5
50K	1	1	Cardiac arrest	0.68	<1
		2	Coma, Cardiac arrest	0.83	<1
		3	Epilepsy, Cardiac arrest, Renal fail.	1.00	7
		4	Shock, Metabolic disord., Coma, Cardiac arrest	1.00	65
	2	1	Cardiac arrest	0.72	<1
		2	Dis. of white blood cells, Cardiac arrest	0.84	<1
		3	Septicemia, Dis. of white blood cells, Cardiac arrest	1.00	6
		4	Resp. fail., Coma, Chronic obs. pulmonary dis., Cardiac arrest	1.00	69
	3	1	Cardiac arrest	0.63	<1
		2	Shock, Secondary malign.	0.83	<1
		3	Secondary malign., Resp. fail., Gastrointestinal hemor.	1.00	7
		4	Diab. mell. with complications, Coma, Cardiac arrest, Renal fail.	1.00	67
	4	1	Cardiac arrest	0.70	<1
		2	Epilepsy, Cardiac arrest	0.83	<1
		3	Nutri. defic., Renal fail., Cardiac arrest	1.00	7
		4	Essential hypertension, Epilepsy, Cardiac arrest, Renal fail.	1.00	64
	5	1	Cardiac arrest	0.62	<1
		2	Chronic obs. pulmonary dis., Cardiac arrest	0.72	<1
		3	Secondary malign., Pneumonia, Acute posthemor. anemia	0.91	7
		4	Septicemia, Metabolic disord., Mainten. chemotherapy, E Codes. effects of medical drugs	1.00	68
100K	1	1	Cardiac arrest	0.69	<1
		2	Coma, Cardiac arrest	0.83	1
		3	Nutri. defic., Cardiac arrest, Acute myocardial infarc.	1.00	14
		4	Urin tract infections, Mental health, Resp. fail., Chemotherapy	1.00	114
	2	1	Cardiac arrest	0.66	<1
		2	Coma, Cardiac arrest	0.83	1
		3	Phlebitis, Chronic ulcer of skin, Cardiac arrest	1.00	13
		4	Mental health, Pulmonary heart dis., Pleurisy, Cardiac arrest	1.00	121
	3	1	Cardiac arrest	0.65	<1
		2	Shock, Second. malign.	0.78	1
		3	Pulmonary heart dis., Dementia, Aspir. pneumonitis	1.00	12
		4	Resp. fail., E Codes. effects of medical drugs, Chronic obstr. pulm. dis. and bronch., Cardiac arrest	1.00	115
	4	1	Cardiac arrest	0.69	<1
		2	Epilepsy, Cardiac arrest	0.80	1
		3	Epilepsy, Complic. of surgical proced., Cardiac arrest	1.00	13
		4	Dis. of white blood cells, Complic. of surgical proced., Cardiac arrest, Bacterial infect.	1.00	116
	5	1	Cardiac arrest	0.63	<1
		2	Coma, Cardiac arrest	0.74	1
		3	Shock, Coma, Cardiac arrest	0.85	12
		4	Septicemia, Connect. tissue dis., Chemotherapy, E Codes. effects of medical drugs	1.00	119