# Exact Approaches for the Maximum Mortality Rate Clique Problem

Parisa Vaghfi Mohebbi
School of Industrial Engineering & Management, Oklahoma State University, Stillwater, Oklahoma, United States, parisa.vaghfi_mohebbi@okstate.edu

Yajun Lu
Department of Management & Marketing, Jacksonville State University, Jacksonville, Alabama, United States, ylu@jsu.edu

Balabhaskar Balasundaram
School of Industrial Engineering & Management, Oklahoma State University, Stillwater, Oklahoma, United States, baski@okstate.edu

This paper develops exact solution methods for the maximum mortality rate clique problem, which aims to identify a cluster of diseases in a comorbidity graph associated with the highest mortality rate among patients whose healthcare encounters are recorded in electronic health records. We establish the NP-hardness of the problem and propose two mixed-integer linear programming formulations, a combinatorial branch-and-bound algorithm, and a logic-based Benders decomposition (LBBD) framework. A hybrid LBBD variant that strategically invokes the combinatorial algorithm to strengthen node relaxations offers a computationally effective approach. The framework accommodates side-constraints on the discovered cliques, such as age-based restrictions, revealing qualitatively distinct lethal disease combinations that are unlikely to be driven solely by elevated baseline mortality in elderly populations. This optimization framework offers a practical and flexible tool for medical researchers seeking to uncover lethal multimorbidities to study progression and potential interactions.

*Keywords*: Multimorbidity analysis; Electronic health record analysis; Comorbidity graph; Logic-based Benders decomposition

## 1. Introduction

Electronic health records (EHRs) capture comprehensive patient information, including demographics, diagnoses, medications, laboratory results, and procedures. The adoption of EHR systems has grown dramatically over the past decade, spurred by initiatives like the U.S. Health Information Technology for Economic and Clinical Health Act of 2009, which provided substantial incentives for certified EHR implementation. By 2015, 84% of hospitals and 87% of office-based physicians had adopted certified systems (Shickel et al. 2018). This widespread adoption has generated vast repositories of structured and unstructured patient data, opening the door to secondary clinical informatics applications beyond primary uses like billing and documentation (Jensen et al. 2012, Yadav et al. 2018).

These rich datasets enable a wide range of statistical and computational techniques for healthcare analytics, including information extraction, representation learning, outcome prediction, phenotyping, de-identification, disease prediction, comorbidity analysis, and patient stratification (Yadav et al. 2018, Sarwar et al. 2022, Jensen et al. 2012). Notable applications include extracting medical concepts from clinical notes (Sarwar et al. 2022), learning patient representations for downstream predictive tasks (Shickel et al. 2018), forecasting clinical outcomes such as hospital readmission or heart failure (Yadav et al. 2018), and automating phenotyping for research cohort selection (Jensen et al. 2012), facilitating cohort-wide investigations and knowledge discovery, ultimately supporting clinical decision-making and personalized medicine.

A stand-out application, from our perspective, among the many analytical opportunities offered by EHRs is comorbidity analysis. The co-occurrence of multiple diseases often produces complex interactions that significantly elevate mortality risk beyond the additive effects of individual conditions (Feinstein 1970). As noted by Redelmeier et al. (1998), a dominant illness can divert clinical attention, potentially leading to the neglect of concurrent conditions, a risk that data-driven methods can help mitigate. Recent studies illustrate the value of mining comorbidity patterns to enable early detection and intervention (Balasubramanian et al. 2024, Mohebbi et al. 2025).

Optimization-based approaches further advance this domain by formulating complex questions as constrained optimization problems. Zhong et al. (2022) cast relational-sequential patient clustering as an optimization problem incorporating custom distance measures that respect demographic hierarchies and diagnosis sequence similarity, yielding clinically meaningful clusters. Vaghfi Mohebbi et al. (2025) introduce *the maximum mortality rate clique problem* to detect lethal cliques in disease comorbidity graphs.

### 1.1. Our Contributions

This article further advances the study of the maximum mortality rate clique problem introduced by Vaghfi Mohebbi et al. (2025). We establish the NP-hardness of the problem and several related variants. We explore two mixed-integer linear programming (MILP) formulations, one based on a product linearization technique from (Vaghfi Mohebbi et al. 2025) and the other based on the Charnes and Cooper (1962) transformation. Direct solution of the two MILP formulations encounters scalability challenges on our test bed as the patient population size grows, confirming that a different algorithmic strategy is needed for

large instances. To address this challenge, we explore multiple complementary approaches that are shown to be far more effective and scalable. First, we develop a new purely combinatorial upper-bounding technique, which is used to transform the enumerative algorithm introduced by Vaghfi Mohebbi et al. (2025) into a branch-and-bound algorithm. Then, we introduce a logic-based Benders reformulation of the problem that avoids the large number of patient-indexed variables present in the MILP formulations (Hooker 2023). The reformulation is naturally suited to a decomposition branch-and-cut algorithm in which optimality and feasibility cuts are generated on demand. Because the mortality rate objective is non-monotone, the optimality cuts take the form of simple no-good cuts, which are inherently weak. We counteract this weakness through two mechanisms: a root node relaxation strengthened by valid inequalities based on the combinatorial upper-bound and a hybrid variant strategy that strategically invokes the combinatorial branch-and-bound algorithm to produce a new valid inequality used to strengthen node relaxations. We conduct an extensive computational study on a real EHR dataset to evaluate our algorithms.

The remainder of this paper is organized as follows. Section 2 formally defines the problem and Section 3 presents the computational complexity results for the relevant decision counterparts. Section 4 introduces two MILP formulations, Section 5 describes a combinatorial branch-and-bound algorithm based on a new upper-bounding technique, and Section 6 develops two variants of a logic-based Benders decomposition approach. Section 7 details the EHR dataset used in this study and reports our computational results. In Section 8, we present a case study to illustrate usefulness of our main contribution: the modeling flexibility and computational effectiveness of our "hybrid logic-based Benders decomposition method". We conclude by discussing our contributions, results, limitations of the study, and outlining directions for future research in Section 9.

## 2. Problem Statement

In formalizing the optimization problem of interest, we assume that EHR data is used to instantiate the inputs to the problem. Let $P$ denote the set of patients represented in the EHRs and the subset of these patients who are assigned an expired status are denoted by $\widetilde{P} \subseteq P$. The set of diseases (or disease categories) denoted by $V$ corresponds to standardized codes based on the International Classification of Diseases (ICD) (Centers for Disease Control and Prevention 2013, 2022). We let $A_u \subseteq P$ denote the patients afflicted

with disease $u \in V$ and $D_i \subseteq V$ denote the diseases afflicting patient $i \in P$. Without loss of generality, we assume that subsets $A_u$ for $u \in V$ and $D_i$ for $i \in P$ are all non-empty.

The same EHR data used to construct the aforementioned inputs to our optimization problem can also be used to build a *comorbidity graph* that captures the co-occurrence of pairs of diseases in the patient population $P$ as edges in an undirected graph. Based on prior studies with comorbidity graphs (Kalgotra et al. 2020, Vaghfi Mohebbi et al. 2025), we use the Ochiai coefficient (akin to cosine similarity for binary vectors) to quantify co-occurrence as $\sigma(u,v) = |A_u \cap A_v|/\sqrt{|A_u| \times |A_v|}$, which normalizes the co-occurrence of a pair of diseases (number of afflicted patients in common) by the geometric mean of the individual occurrences in $P$ (Ochiai 1957, Salton and McGill 1983). Then, with a user-specified threshold $\Delta \in [0,1]$, we can define the edge set of our comorbidity graph $G = (V, E)$ as $E := \left\{\{u,v\} \in \binom{V}{2} : \sigma(u,v) \geqslant \Delta\right\}$, where $\binom{V}{2} := \{\{u,v\} : u, v \in V, u \neq v\}$.

Definition 1. Given a subset of diseases $C \subseteq V$, its mortality rate $\mu(C)$ is defined as follows: $\mu(C) := |\operatorname{num}(C)|/|\operatorname{den}(C)|$, where $\operatorname{den}(C) := \bigcap_{u \in C} A_u$ and $\operatorname{num}(C) := \operatorname{den}(C) \cap \widetilde{P}$. The denominator of the mortality rate is the size of the set of patients in $P$ who are afflicted with all of the diseases in $C$, while the numerator is the number of such patients who have expired. We take $\mu(C)$ to be zero if either $C$ or $\operatorname{den}(C)$ is empty. A *clique* (pairwise adjacent vertices) is used to model a tightly-knit cluster of co-occurring diseases.

Definition 2. Given a positive integer $\ell$, we say that a clique $C$ is *$\ell$-frequent* if $|\operatorname{den}(C)| \geqslant \ell$ and it is a non-$\ell$-frequent clique otherwise. We say that $C$ is a *minimal* non-$\ell$-frequent clique if $|\operatorname{den}(C)| \leqslant \ell - 1$ and $|\operatorname{den}(C \backslash \{u\})| \geqslant \ell$ for every $u \in C$.

The *maximum mortality rate clique problem* (MMRCP) can be stated as $\max\{\mu(C) : C$ is a non-empty $\ell$-frequent clique$\}$. The precise interpretation of $\mu(C)$ depends on what the EHR dataset characterizes as mortality status (expired, hospice care, time frame considered) for the patients included in $\widetilde{P}$, *e.g.*, it could indicate that the patient expired or was discharged to hospice care inside a 1-year time window after their last encounter.

The version of the MMRCP introduced in (Vaghfi Mohebbi et al. 2025), seeks to find an $\ell$-frequent clique *of size at most $b$* in the comorbidity graph $G$ that maximizes the mortality rate, especially given some pre-existing members of the clique. Focusing on small but lethal cliques is of much value to a clinician who can use the information to proactively decide care when assessing future mortality risk of a patient with multiple comorbidities. The size restriction also enabled the development of effective enumerative approaches. We

consider the question from a medical researcher's perspective for whom high mortality rate cliques of small sizes may not provide adequate information when they wish to better understand the role all co-occurring conditions might play, their progression over time, and interactions. We therefore focus on the size-unrestricted counterpart in this study, beginning with their computational complexity.

## 3. Complexity

In this section, we establish the NP-completeness of three decision problems related to our optimization problem of interest, which is to find a non-empty $\ell$-frequent clique in the comorbidity graph $G$ with maximum mortality rate. In the first two results we use reductions from the classical 3-Satisfiability problem also stated below.

**Problem**: High Mortality Rate Clique (Hmrc)
**Instance**: A comorbidity graph $G=(V,E)$, patient set $P$, deceased patient subset $\widetilde{P}\subseteq P$, incidence sets $A_u\subseteq P$ for each $u\in V$, positive integer $\ell$, and a rational number $\bar{\mu}\in(0,1]$.
**Question**: Does $G$ contain an $\ell$-frequent clique $C$ with $\mu(C)\geqslant\bar{\mu}$?

**Problem**: 3-Satisfiability (3Sat)
**Instance**: A Boolean formula in conjunctive normal form $\mathcal{F}=\bigwedge_{i=1}^{m}(f_{i1}\vee f_{i2}\vee f_{i3})$, and each literal $f_{ij}\in\{x_k,\neg x_k\}$ for some $k=1,2,\ldots,n$.
**Question**: Is there an assignment for Boolean variables $(x_1,x_2,\ldots,x_n)$ that satisfies $\mathcal{F}$?

Theorem 1. Hmrc *is NP-complete for every positive integer* $\ell$.

*Proof.* Construct the Hmrc instance from the 3SAT instance as follows. Let $V:=\bigcup_{i=1}^{m}\{v_{i1},v_{i2},v_{i3}\}$, associating every literal in every clause to a corresponding vertex. For each pair $\{u,v\}\in\binom{V}{2}$ an edge is added to $E$ if and only if $u$ and $v$ correspond to literals that are not negations of each other and belong to distinct clauses. The deceased patient set includes $\ell$ patients denoted by $\widetilde{P}:=\{q_1,\ldots,q_\ell\}$ and the live patient set includes one patient corresponding to each clause, $P\backslash\widetilde{P}:=\{p_1,\ldots,p_m\}$. We let the set of diseases afflicting an arbitrary live patient $p_i$ be $D_{p_i}:=V\backslash\{v_{i1},v_{i2},v_{i3}\}$, i.e., all diseases except those corresponding to that patient's clause $i$. For each deceased patient $q_i$ we let $D_{q_i}:=V$. Then, the set of patients afflicted by an arbitrary disease $v_{ij}$ is given by $A_{v_{ij}}=P\backslash\{p_i\}$.

Finally, we set $\bar{\mu} = 1$, which completes our reduction, easily verified to be a polynomial-time transformation. For any $C \subseteq V$, we say $C$ does not cover clause $i$ if $C \cap \{v_{i1}, v_{i2}, v_{i3}\} = \varnothing$. Observe that for any nonempty $C \subseteq V$,

$$\begin{aligned}
\operatorname{den}(C) &= \bigcap_{v_{ij} \in C} P \backslash \{p_i\} = \widetilde{P} \cup \{p_i \mid \text{clause } i \text{ is not covered by } C\}, \\
\operatorname{num}(C) &= \widetilde{P} \cap \operatorname{den}(C) = \widetilde{P}, \text{ and} \\
\mu(C) &= \frac{\ell}{\ell + k}, \text{ where } k \text{ is the number of clauses not covered by } C.
\end{aligned}$$

Now, we show that $\mathcal{F}$ is satisfiable if and only if the HMRC instance is a yes-instance. ($\Rightarrow$) Given a satisfying assignment for $x$, include in set $C$, one vertex corresponding to a literal $f_{ij}$ that is true in each clause $i$. That is, $C := \bigcup_{i=1}^{m} \{v_{ij} : f_{ij} = \texttt{true}, j \text{ is a minimum}\}$ and $|C| = m$. If we consider a pair of distinct vertices from $C$, they must correspond to distinct clauses and, as they both correspond to true literals, they cannot represent a Boolean variable and its negation. As every clause is covered, $k = 0$, and $|\operatorname{den}(C)| = \ell$. Hence, $C$ is an $\ell$-frequent clique with $\mu(C) = \bar{\mu} = 1$, implying that HMRC instance is a yes-instance. ($\Leftarrow$) Suppose $C \subseteq V$ is an $\ell$-frequent clique for which $\mu(C) \geqslant \bar{\mu}$. Then, we have $\ell/(\ell + k) \geqslant 1 \implies k = 0$. Hence, all clauses are covered and the literals corresponding to vertices in $C$ can all be set to one, without conflicts, to obtain an assignment that satisfies $\mathcal{F}$. (Any $x_k$ not fixed in this manner can be set to any assignment.)

It is easy to verify that HMRC is in class NP and therefore it is NP-complete. □

REMARK 1. The reduction shows that HMRC is NP-complete if $\ell$ is an arbitrary positive integer specified in the input or if it is a constant fixed in the problem. In particular, HMRC is NP-complete even if $\ell = 1$. It also follows from our reduction that HMRC remains NP-complete if we seek a clique of size at most $b$, exactly $b$ or at least $b$, as the reduction extends to these variants by choosing the input parameter $b = m$, the number of clauses.

Suppose we are only interested in cliques with mortality rate at least 0.9, then Theorem 1 is no longer applicable as $\bar{\mu}$ is fixed in the problem and not arbitrary (instantiated in the input). The reduction used in Theorem 1 also implies HMRC is NP-complete if $\bar{\mu}$ is the constant one, but not for a different constant. Theorem 2 that follows shows that HMRC remains NP-complete even if we fix $\bar{\mu}$ to be any rational constant in the interval $(0, 1)$.

THEOREM 2. HMRC *is NP-complete for every rational* $\bar{\mu} \in (0, 1)$.

*Proof.* The construction of $G$ in the HMRC instance is identical to the proof of Theorem 1. The live patient set includes a new group of $t$ patients denoted by $\{r_1, \ldots, r_t\}$, where $t := \left\lfloor \frac{m(1-\bar{\mu})}{\bar{\mu}} \right\rfloor$. With the deceased patient set $\widetilde{P} := \{q_1, \ldots, q_m\}$, the live patient set is given by $P \backslash \widetilde{P} = \{p_1, \ldots, p_m\} \cup \{r_1, \ldots, r_t\}$. The disease incidence sets are defined as follows: $D_{p_i} := V \backslash \{v_{i1}, v_{i2}, v_{i3}\}$, $D_{r_i} := V$, and $D_{q_i} := V$. As before, the set of patients afflicted by an arbitrary disease $v_{ij} \in V$ is given by $A_{v_{ij}} = P \backslash \{p_i\}$. Finally, we set $\ell = m + t$ completing our reduction. For any nonempty $C \subseteq V$,

$$\begin{aligned}
\operatorname{den}(C) &= \bigcap_{v_{ij} \in C} P \backslash \{p_i\} = \widetilde{P} \cup \{r_1, \ldots, r_t\} \cup \{p_i \mid \text{clause } i \text{ is not covered by } C\}, \\
\operatorname{num}(C) &= \widetilde{P} \cap \operatorname{den}(C) = \widetilde{P}, \text{ and} \\
\mu(C) &= \frac{m}{m + t + k}, \text{ where } k \text{ is the number of clauses not covered by } C.
\end{aligned}$$

Now, we show that $\mathcal{F}$ is satisfiable if and only if the HMRC instance is a yes-instance. ($\Rightarrow$) Given a satisfying assignment for $x$, the set $C := \bigcup_{i=1}^{m} \{v_{ij} : f_{ij} = \texttt{true}, j \text{ is a minimum}\}$ is a clique of size $m$. As every clause is covered, $k = 0$, and $|\operatorname{den}(C)| = m + t = \ell$. Hence, $C$ is an $\ell$-frequent clique with,

$$\mu(C) = \frac{m}{m + t} \geqslant \frac{m}{m + m(1 - \bar{\mu})/\bar{\mu}} = \bar{\mu}.$$

($\Leftarrow$) Suppose $C \subseteq V$ is an $\ell$-frequent clique for which $\mu(C) \geqslant \bar{\mu}$. Then, we have,

$$\mu(C) = \frac{m}{m + t + k} \geqslant \bar{\mu} \iff k \leqslant \frac{m(1 - \bar{\mu})}{\bar{\mu}} - \left\lfloor \frac{m(1 - \bar{\mu})}{\bar{\mu}} \right\rfloor < 1,$$

which implies $k = 0$. Hence, all clauses are covered and like before we can obtain a truth assignment for $\mathcal{F}$. It is easy to verify that HMRC is in class NP. □

Next, we show the intractability of a simple feasibility question that seeks any subset of $b$ diseases that are $\ell$-frequent, with no clique or mortality rate requirements.

**Problem**: $\ell$-FREQUENT DISEASE SUBSET (FDS)
**Instance**: A set of diseases $V$, patient set $P$, incidence sets $A_u \subseteq P$ for each $u \in V$, and positive integers $\ell$ and $b$.
**Question**: Is there a $C \subseteq V$ such that $|C| \geqslant b$ and $|\operatorname{den}(C)| \geqslant \ell$?

**Problem**: BALANCED BICLIQUE IN BIPARTITE GRAPH (B3G)
**Instance**: Bipartite graph $G = (L \cup R, E)$ and a positive integer $t$.
**Question**: Does $G$ contain a complete bipartite subgraph with equal partitions of size $t$?

THEOREM 3. FDS *is NP-complete.*

*Proof.* The reduction is from B3G. In constructing the FDS instance, we set $V := L$, $P := R$, and $\ell = b = t$. Denoting the neighborhood set of a vertex $u \in L \cup R$ in $G$ by $N(u)$, we can define $A_u := N(u)$ for each $u \in V$. Then, $D_i$ for each $i \in P$ is given by $N(i)$. Note that for any $C \subseteq V$, we have $\text{den}(C) = \bigcap_{u \in C} N(u)$.

If $C \subseteq L$ and $B \subseteq R$ induces a balanced biclique in $G$ with $t$ vertices in each partition, then we have $|C| = b$, $B \subseteq \text{den}(C)$, and $|\text{den}(C)| \geqslant |B| = \ell$. Conversely, if $C$ is a set of at least $b$ diseases with $\text{den}(C) = \bigcap_{u \in C} N(u)$ containing at least $\ell$ patients, then $C \subseteq L$ and $\text{den}(C) \subseteq R$, once restricted to $t$ vertices in each, corresponds to a balanced biclique. □

The scalability expectations of exact algorithms for MMRCP are tempered by the foregoing hardness results, which show that many aspects of the problem make it challenging to solve. Nonetheless, this problem needs to be solved on large datasets derived from EHR databases, so, we focus next on advancing the methodological toolkit for this problem.

## 4. Integer Programming Formulations

Formulation (1) of the MMRCP, is a mixed-integer nonlinear fractional program that uses a vector of binary variables $x$ to indicate vertices included in a clique and a vector of binary variables $y$ to identify patients afflicted by all diseases included in the clique. We let $\bar{G} = (V, \bar{E})$ denote the complement graph of the comorbidity graph $G = (V, E)$, and we use the short form $uv$ to denote an edge $\{u, v\}$. The set $\mathcal{X} \subseteq \{0,1\}^{|V|}$ denotes the characteristic vectors of *non-empty* cliques in $G$. The fractional objective function (1a) represents the mortality rate being maximized. Constraint (1b) forces $y_i = 0$ whenever a disease $u \notin D_i$ is selected. If the selected clique is a subset of $D_i$, then constraint (1c) forces $y_i = 1$. Together, these constraints ensure that $y_i = 1$ if and only if patient $i$ has every disease in the selected clique. Constraint (1d) ensures that we only consider $\ell$-frequent cliques.

$$\max \frac{\sum_{i \in \tilde{P}} y_i}{\sum_{i \in P} y_i} \tag{1a}$$

$$s.t. \quad y_i \leqslant 1 - x_u \qquad \forall u \in V \backslash D_i,\ i \in P \tag{1b}$$

$$y_i \geqslant 1 - \textstyle\sum_{u \in V \backslash D_i} x_u \quad \forall i \in P \tag{1c}$$

$$\textstyle\sum_{i \in P} y_i \geqslant \ell \tag{1d}$$

$$x \in \mathcal{X} \text{ and } y \in \{0,1\}^{|P|} \tag{1e}$$

Vaghfi Mohebbi et al. (2025) obtained an equivalent MILP by linearizing the fractional objective and we additionally explore a Charnes and Cooper (1962) style linearization as well in this article. Both MILP formulations are provided in Appendix A.

## 5. Combinatorial Branch-and-Bound Algorithm

An algorithm that enumerates cliques to find a list of lethal cliques (maximum mortality rate cliques included) was introduced by Vaghfi Mohebbi et al. (2025). This algorithm adapted the classical Bron and Kerbosch (1973) algorithm to enumerate all (not only maximal) cliques while pruning the search when a non-$\ell$-frequent clique is encountered. The approach proved to be effective on very large-scale instances to find $\ell$-frequent cliques of size at most $b$ (usually $b \leqslant 4$). We improve this algorithm by developing an upper-bounding scheme that enables additional pruning in a depth-first search (DFS) order combinatorial branch-and-bound (CBB) algorithm.

We derive an upper-bound on the mortality rate $\mu(C)$ of an $\ell$-frequent clique $C$, assuming that $C \supset C'$. Note that if $C$ is an $\ell$-frequent clique, then so is $C'$. Clearly, $|\operatorname{num}(C)| \leqslant |\operatorname{num}(C')|$ and $|\operatorname{den}(C)| \geqslant \ell$. Therefore,

$$\mu(C) \leqslant \gamma(C') := \frac{|\operatorname{num}(C')|}{\ell}, \tag{2}$$

an upper-bound that could be used if we know that vertices in $C'$ are included in the solution. The "$\gamma$-upper-bound" is monotonically non-increasing, i.e., $\gamma(C'') \leqslant \gamma(C')$ for $C \supset C'' \supset C'$, making it a useful pruning tool in a DFS-order search where the current clique monotonically expands as we traverse down the search tree.

Algorithm 2 in Appendix B searches the graph in DFS order and prunes the search based on both $\gamma$-upper-bounds and violation of $\ell$-frequency. The algorithm makes recursive calls to itself, while expanding the current clique $C$. The function calls to `DfsBB` also include the set $L$ of candidate vertices (common neighbors of vertices in $C$), the denominator and numerator terms associated with $C$. The reason for passing the last two arguments is to allow for their incremental updates as $C$ adds on another vertex. In each iteration of the for-loop we attempt to add a vertex $v \in L$ to enlarge current clique $C$. We ensure that the new clique $C \cup \{v\}$ is $\ell$-frequent and $\gamma(C)$ is larger than the best objective value currently known before making the next recursive call. Prior to making a recursive call the incumbent solution and objective are updated, if necessary, based on the new solution.

## 6. Logic-Based Benders Reformulation

Both MILP formulations presented in Appendix A rely on linearizing the fractional objective function, introducing additional variables and constraints in the process. Although $|V|$ is moderately sized when representing disease categories based on ICD codes, most EHR data contain information for an extremely large population of patients. The size of $P$ depends on the EHR source and time period covered, and is usually massive. The $y$-variables that indicate patients afflicted by all the diseases selected by $x \in \mathcal{X}$ are used in the fractional integer program (1), as the calculation of the mortality rate objective function is based on the deceased and live patient count. Just the constraints (1b)-(1d), which couple $x$ and $y$ variables and ensure that the clique selected is $\ell$-frequent, correspond to $2|P| + 3\sum_{i\in P}|V\backslash D_i|$ non-zero constraint coefficients. Assuming $\min_{i\in P}|V\backslash D_i| > \epsilon|V|$, the aforementioned constraints alone add $(2 + 3\epsilon|V|)|P|$ non-zero constraint coefficients (and usually $\epsilon > 0.5$). These constraints are also used in MILP (13) in Appendix A and their "continuous counterparts" are used in the Charnes–Cooper reformulation (14) in Appendix A. The growth in size of the formulation with $|P|$ poses a key computational challenge as observed by Vaghfi Mohebbi et al. (2025) when solving instances with over 10 million patients using a lazy-cut implementation of MILP (13).

An important contribution of this study is the alternate approach we developed to avoid the $y$-variables and the linearization of the fractional objective function. We introduce a *logic-based Benders reformulation* of the problem that is naturally designed to be decomposed (Hooker 2023), a characteristic neither of the MILP formulations have. Logic-based Benders decomposition (LBBD) offers a general framework for reformulating, decomposing, and solving challenging combinatorial optimization problems and is the basis for the combinatorial Benders cuts introduced by Codato and Fischetti (2006) for reformulating and solving MILP formulations with big-M constraints used to model conditional implications (constraints "activated" by binary variables).

Notable applications of LBBD techniques can be found in a variety of domains including avionics scheduling with partial assignment acceleration (Karlsson and Rönnberg 2022), preemptive flexible job-shop scheduling (Juvin et al. 2023), stochastic distributed operating room scheduling with binary decision diagrams (Guo et al. 2021), chronic outpatient scheduling in answer set programming (Cappanera et al. 2023), multi-manned assembly

line balancing with symmetry-breaking cuts (Michels et al. 2019), and two-dimensional bin packing using area-based models and no-good cuts (Côté et al. 2021).

A fundamental challenge to applying LBBD to our problem stems from the non-monotone nature of the mortality rate, as a function of the clique of selected diseases (Vaghfi Mohebbi et al. 2025). This necessitates the use of the so-called *simple no-good* optimality cuts that are extremely weak (Hooker 2023). Improving our logic-based Benders reformulation, especially its use inside a decomposition branch-and-cut (DBC) algorithm is an important focus of this paper. The feasibility cuts, by contrast, are cover inequalities based on minimal non-$\ell$-frequent cliques, which are much stronger.

We present the reformulation next, using the $x$-variables indicating the clique and a scalar variable $z$ designed to capture the mortality rate. We require a few additional notations introduced next. Given binary vectors $x, \widetilde{x} \in \{0,1\}^{|V|}$, the Hamming distance between them, denoted by $\mathcal{H}(x, \widetilde{x})$, is the number of components in which $x$ and $\widetilde{x}$ differ. This can be expressed as: $\mathcal{H}(x, \widetilde{x}) = \sum_{u \in J_0(\widetilde{x})} x_u + \sum_{u \in J_1(\widetilde{x})} (1 - x_u)$, where $J_0(\widetilde{x}) \coloneqq \{u \in V : \widetilde{x}_u = 0\}$ and $J_1(\widetilde{x}) \coloneqq \{u \in V : \widetilde{x}_u = 1\}$. For convenience and to avoid cluttered notation, we "overload" the mortality rate $\mu(\cdot)$ and $\gamma$-upper-bound notations as follows: $\mu(x) \coloneqq \mu(J_1(x))$, $\mu(uv) \coloneqq \mu(\{u, v\})$, $\gamma(uv) \coloneqq \gamma(\{u, v\})$, $\mu(u) \coloneqq \mu(\{u\})$, and $\gamma(u) \coloneqq \gamma(\{u\})$. The logic-based Benders reformulation of the maximum mortality rate problem is as follows.

$$\max\ z \tag{3a}$$

$$s.t. \quad z \leqslant \mu(\tilde{x}) + (1 - \mu(\tilde{x}))\, \mathcal{H}(x, \tilde{x}) \quad \forall\, \widetilde{x} \in \mathcal{X}_o \tag{3b}$$

$$\sum_{u \in J_1(\widetilde{x})} x_u \leqslant |J_1(\widetilde{x})| - 1 \qquad \forall\, \widetilde{x} \in \mathcal{X}_f \tag{3c}$$

$$x \in \mathcal{X} \tag{3d}$$

where, $\mathcal{X}_o$ and $\mathcal{X}_f$ partition $\mathcal{X}$ such that $\mathcal{X}_o$ contains incidence vectors of $\ell$-frequent cliques and $\mathcal{X}_f$ contains incidence vectors of minimal non-$\ell$-frequent cliques.

Constraint (3b) is the simple no-good optimality cut that forces $z \leqslant \mu(x)$ when $x = \widetilde{x}$, yielding a correct computation of the objective function value for the feasible $x$. It is made redundant for every $x \neq \widetilde{x}$. Hence, every binary vector $x \neq \widetilde{x}$ satisfies this constraint for all $z \in [0, 1]$. Constraint (3c) is a cover inequality that serves as a feasibility cut to eliminate cliques that are not $\ell$-frequent, using the fact that every superset of a non-$\ell$-frequent clique is also not $\ell$-frequent. To effectively use this formulation in a DBC algorithm we need to address several key issues such as the master problem used at the root node, cuts added to strengthen the relaxation, as well as other improvements. We discuss these details next.

### 6.1. Root Relaxation and Delayed Constraint Generation

The optimality cuts (3b) and feasibility cuts (3c) in the logic-based Benders reformulation (3) can be exponentially large in the worst case as they are in correspondence with non-empty cliques in the comorbidity graph. As is typically the case in a Benders decomposition algorithm, these cuts are generated and added in a delayed fashion when the optimal solution to a relaxation violates one of these cuts, i.e., a delayed row generation scheme. Because an initial relaxation that does not include all (or any) of the optimality or feasibility cuts is an integer program, which is solved with a branch-and-cut (BC) algorithm, these cuts may be added at the nodes of the BC tree when an integral $x$ is encountered that violates one of them. This approach is usually referred to as a decomposition-branch-and-cut (DBC) algorithm, and that is how we intend to implement this formulation.

The strength of the relaxation at the root node of the BC tree is an important factor in the overall performance of the DBC algorithm, especially since our optimality cuts are extremely weak. For that same reason, initializing the root node relaxation with constraints corresponding to a subset of $\mathcal{X}$ may not necessarily be favorable either. So, we develop valid inequalities that are intended for use in the root node relaxation. To this end, we use the $\gamma$-upper-bounds for a single vertex $v$ or the endpoints of an edge $uv$.

Our root node relaxation for the DBC algorithm, Formulation (4), combines valid inequalities based on the $\gamma$-upper-bounds when they are non-trivial (i.e., less than one) and the special case of the simple no-good constraints (3b) otherwise. Let $\mu_{LB} \coloneqq \max_{u\in V} \mu(u)$ and $e_u$ denotes the $|V|$-dimensional unit vector with one in position $u$.

$$\max\ z \tag{4a}$$

$$s.t.\quad z \leqslant \mu(v) + (1-\mu(v))\,\mathcal{H}(x, e_v) \qquad \forall\, v\in V : \gamma_v \geqslant 1 \tag{4b}$$

$$z \leqslant \gamma_v x_v + (1-x_v) \qquad \forall\, v\in V : \gamma_v < 1 \tag{4c}$$

$$z \leqslant \mu(uv) + (1-\mu(uv))\,\mathcal{H}(x, e_u+e_v) \qquad \forall\, uv\in E : \gamma(uv) \geqslant 1 \tag{4d}$$

$$z \leqslant \gamma(uv)(x_u+x_v-1) + (2-x_u-x_v) \qquad \forall\, uv\in E : \gamma(uv) < 1 \tag{4e}$$

$$z \geqslant \mu_{LB} \tag{4f}$$

$$x \in \mathcal{X} \tag{4g}$$

The DBC algorithm begins solving Formulation (4) at the root node. At some node of the BC search tree, if $\widetilde{x}$ is integral in the optimal solution to the linear programming (LP)

relaxation that is found, we must verify if it is feasible to Formulation (3) using the separation procedure described in Algorithm 3 in Appendix B. As explained in Remark 2 that follows, $J_1(\widetilde{x})$ will be a clique. If it is not $\ell$-frequent we add a feasibility cut; otherwise, we add a violated optimality cut, if one exists. If a violated optimality or feasibility cut exists, one can be found in polynomial time and returned to the BC node. The node LP relaxation is re-solved if a violated cut is returned, otherwise it is pruned by feasibility. Algorithm 4 in Appendix B describes the procedure to make a non-$\ell$-frequent clique minimal, in order to add feasibility cuts based on minimal covers.

REMARK 2. Suppose the comorbidity graph $G = (V, E)$ has $k$ connected components $H_1, \ldots, H_k$. Consider the LP relaxation of $\mathcal{X}$ denoted by $\mathcal{X}_{LP} = \{x : (x, f) \text{ satisfies } (5)-(9)\}$, assuming $\mathcal{X}$ is described by a standard approach using binary variables $x_u : u \in V$ and binary component variables $f_i : i = 1, \ldots, k$.

$$x_u + x_v \leqslant 1 \qquad \forall\, uv \in E(\bar{H}_i),\ i = 1, \ldots, k \tag{5}$$

$$x_u \leqslant f_i \qquad \forall\, u \in V(H_i),\ i = 1, \ldots, k \tag{6}$$

$$\textstyle\sum_{i=1}^{k} f_i = 1 \tag{7}$$

$$\textstyle\sum_{u \in V} x_u \geqslant 1 \tag{8}$$

$$(x, f) \in [0,1]^{|V|} \times [0,1]^k \tag{9}$$

If $x^* \in \mathcal{X}_{LP}$, then the set $C := \{v \in V : x^*_v > \frac{1}{2}\}$ is a clique in $G$. Suppose, for contradiction, that $C$ contains non-adjacent vertices $u$ and $v$. If $u \in V(H_i)$ and $v \in V(H_j)$, constraints (6) imply $f^*_i \geqslant x^*_u > \frac{1}{2}$ and $f^*_j \geqslant x^*_v > \frac{1}{2}$. Hence, either constraint (5) or constraint (7) is violated.

### 6.2. A Hybrid LBBD Approach

This section describes hybrid strategies that use the combinatorial branch-and-bound algorithm introduced in Section 5 to generate strong valid inequalities incorporated into the logic-based Benders DBC algorithm. Theorem 4 that follows, establishes a valid inequality that makes such a hybrid approach possible.

THEOREM 4 **(CBB cut)**. *Consider an $\ell$-frequent clique $\widetilde{C}$, and let $\bar{C}$ denote an $\ell$-frequent clique containing $\widetilde{C}$ with maximum mortality rate. If $x$ corresponds to an arbitrary $\ell$-frequent clique in $G$ and $z = \mu(x)$, then $(x, z)$ satisfies the following inequality:*

$$z \leqslant \mu(\bar{C}) + (1 - \mu(\bar{C})) \sum_{u \in \widetilde{C}} (1 - x_u). \tag{10}$$

*Proof.* If $\widetilde{C} \subseteq J_1(x)$, then $\mu(x) \leqslant \mu(\bar{C})$ by the definition of $\bar{C}$, as $x$ corresponds to an $\ell$-frequent clique containing $\widetilde{C}$ in $G$. Otherwise, $\widetilde{C} \backslash J_1(x) \neq \varnothing$ and inequality (10) is trivially satisfied because, $\sum_{u \in \widetilde{C}} (1 - x_u) \geqslant 1$ and the right-hand side is at least 1. □

At nodes of the BC tree when an integral solution is encountered, we add the CBB cut whenever the number of vertices included in the clique is at least a user-specified threshold $\tau$. The motivation for the hybrid approach is that the LBBD benefits from the CBB algorithm while also retaining the expressiveness of the underlying integer programming formulation. In particular, if $\mathcal{X}$ included "side constraints" that we discuss in Section 8, we could still use this approach to produce valid cuts without modifying the CBB algorithm. The rationale behind the separation strategy proposed in Algorithm 1 is based on the following observations. First, it makes the computationally expensive CBB call only when we already have a "promising" $\ell$-frequent clique of reasonable size, i.e., an integral optimum to the relaxation with at least $\tau$ vertices. CBB can now work on a reduced graph induced by $J_1(\widetilde{x}) \cup \bigcap_{u \in J_1(\widetilde{x})} N(u)$ and potentially produce a strong valid inequality in a reasonable amount of time. Second, because the call is made only on integral solutions that violate the simple no-good cut, the number of such CBB calls is controlled.

A second way we use the CBB algorithm is to strengthen the root node relaxation. We generate CBB cuts (10) for some user-specified number of vertices with the highest individual mortality rates and add it to the root node relaxation (4). Based on these calls to the CBB algorithm, we can potentially improve $z_{LB}$ used in constraint (4f) using the highest mortality rate clique encountered. We refer to the resulting algorithm henceforth as the "hybrid LBBD" algorithm.

## 7. Computational Study

To evaluate the performance of the proposed algorithms, we conduct a comprehensive set of experiments on test instances sampled from a real EHR dataset with 223,291 patients of varying patient cohort sizes ranging from 1,000 to 30,000 patients. We compare the algorithmic approaches considered in this paper such as exact combinatorial algorithms, solving MILP formulations with some enhancements, and variants of the LBBD algorithm.

The discussion of our computational study is organized as follows. We describe our test bed and its development including the construction of the comorbidity graph in Section 7.1. Experimental settings and other common details are covered in Section 7.2. In Section 7.3,

**Algorithm 1:** Hybrid Logic-Based Benders Decomposition: Separation Procedure

**Input:** $\widetilde{x} \in \mathcal{X}$, $\widetilde{z} \in [0,1]$, and positive integer $\tau$

**Output:** Violated optimality, CBB, or feasibility cut, if detected

**if** $J_1(\widetilde{x})$ *is not* $\ell$*-frequent* **then**
  $\widetilde{C} \leftarrow$ `Minimalize`$(J_1(\widetilde{x}))$
  **return** violated feasibility cut: $\sum_{u \in \widetilde{C}} x_u \leqslant |\widetilde{C}| - 1$
**else if** $\widetilde{z} > \mu(\widetilde{x})$ *and* $|J_1(\widetilde{x})| < \tau$ **then**
  **return** violated optimality cut: $z \leqslant \mu(\widetilde{x}) + (1 - \mu(\widetilde{x})) \times \mathcal{H}(x, \widetilde{x})$
**else if** $\widetilde{z} > \mu(\widetilde{x})$ *and* $|J_1(\widetilde{x})| \geqslant \tau$ **then**
  $\bar{C}, \bar{\mu} \leftarrow$ Output of Algorithm 2 called with $C^0 = J_1(\widetilde{x})$
  **return** violated optimality cut: $z \leqslant \mu(\widetilde{x}) + (1 - \mu(\widetilde{x})) \times \mathcal{H}(x, \widetilde{x})$
  and CBB cut: $z \leqslant \bar{\mu} + (1 - \bar{\mu}) \sum_{u \in J_1(\widetilde{x})} (1 - x_u)$
**else**
  **return** no violated cuts found

we compare the modified Bron–Kerbosch algorithm and the CBB algorithm. We evaluate MILP formulations in Section 7.4 and assess variants of the LBBD approach in Section 7.5.

### 7.1. Testbed Preparation

We use version 3.1 of the Medical Information Mart for Intensive Care IV (MIMIC-IV) (Johnson et al. 2023), a publicly available electronic health record database derived from admissions at Beth Israel Deaconess Medical Center (BIDMC) in collaboration with the Massachusetts Institute of Technology. MIMIC-IV spans 2008–2019 and includes structured data on vital signs, laboratory measurements, medications, procedures, diagnoses codes, and de-identified clinical notes. All data originate from routine clinical care, undergo de-identification, and are released only to credentialed researchers who complete human subjects training and sign a data use agreement. The BIDMC Institutional Review Board waived informed consent for sharing this resource. Full details on data acquisition, transformation, and de-identification appear in (Johnson et al. 2023).

MIMIC-IV is accessible via PhysioNet (Goldberger et al. 2000), where the `hosp` and `icu` modules are hosted under the MIMIC-IV project. In this study, unique patient identifiers

and in-hospital mortality status are extracted from `admissions.csv` using the `subject_id` and `hospital_expire_flag` columns, respectively. The `hospital_expire_flag` indicates whether a patient died during the hospital stay. For patients who expired outside the hospital, mortality information is obtained from `patients.csv` via the `dod` column, which records the patient's date of death.

All ICD-coded diagnoses for each patient are obtained from `diagnoses_icd.csv` located in the `hosp` module. Following the MIMIC-IV de-identification protocol, dates of death are censored at one year after the patient's last hospital discharge; a null `dod` therefore indicates that the patient was known to be alive at least one year beyond their final encounter, while a non-null `dod` indicates death within that one-year window. No inferences about mortality beyond one year can be drawn. Consequently, the overall mortality rate in the dataset is 16.51%, representing the one-year post-discharge mortality across the full cohort of unique patients. This definition ensures that the mortality rate is comparable across patients whose last encounter occurred at any point in the 2008–2019 study window.

The diagnoses in the MIMIC-IV dataset are recorded using both the International Classification of Diseases, 9th Revision (ICD-9) and 10th Revision (ICD-10). This dual-coding scheme introduces redundancy, as the same clinical condition may be represented by distinct codes across systems. Furthermore, because the coding framework assigns distinct codes to closely related conditions that differ only by location, cause, or clinical circumstance, it can give the impression that patients have more comorbidities than they really do. To mitigate these complications, we map all ICD-9 and ICD-10 diagnosis codes to clinically coherent, higher-level categories using the Clinical Classifications Software (CCS) for ICD-9 and its successor, CCS Refined (CCSR) for ICD-10 (Agency for Healthcare Research and Quality 2021). These tools, developed by the Healthcare Cost and Utilization Project (HCUP), aggregate ICD codes into a manageable set of mutually exclusive, medically meaningful disease groups. This reduces duplication, improves comparability across coding eras, and improves interpretability (Malecki et al. 2024).

The MIMIC-IV database, from which we sample and create test instances of different sizes, contains 223,291 patients in total, of whom 36,872 are expired, yielding an overall mortality rate of 16.51%. The number of diseases per patient ranges from 1 to 194, with an average of 14.12. From the full set of 223,291 unique patients, we created seven cohort sizes, respectively with 1,000, 5,000, 10,000, 15,000, 20,000, 25,000, and 30,000 patients. For each

cohort size $p$, we drew five independent simple random samples without replacement by selecting $p$ patients uniformly at random from the complete cohort (each patient having equal probability 1/223,291). To ensure exact reproducibility while adding variety, the five replicates for each cohort size were generated using one of the fixed random seeds 13, 18, 35, 66, 74, 100, and 1318. The resulting 35 sampled datasets constitute the test bed used in all subsequent experiments.

In the comorbidity graph model introduced in Section 2, each high-level disease is associated with a vertex and an edge between vertices $u$ and $v$ is added if the Ochiai coefficient $\sigma(u,v) \geqslant \Delta$. To select an appropriate threshold $\Delta$, we follow the methodology described in (Kalgotra et al. 2020). As the statistical significance of the Ochiai coefficient cannot be computed directly, Kalgotra et al. (2020) suggest using the $\phi$-correlation coefficient as a proxy. This allows us to estimate the number of edges that are significant at our chosen significance level $\alpha = 0.01$. Given our large sample size, we adopt this significance level to ensure robustness. Using this procedure, we identify 8,539 statistically significant edges based on the $\phi$-correlation coefficient, which corresponds to a threshold $\Delta = 0.1191$. We therefore select this value of $\Delta$ as the cutoff for edge inclusion. The resulting comorbidity graph contains 467 vertices and 8,539 edges (8% edge density), consisting of a large connected component with 465 vertices and another with 2 vertices.

### 7.2. Experimental Settings and Other Details

All experiments were conducted on a large memory high-performance compute node with an Intel Xeon Gold 6130F CPU @ 2.10 GHz and 1.5 TB RAM, using Gurobi 12.0.0 as the MILP solver. A time limit of 7,200 seconds (2 hours) was imposed for all methods. Delayed addition of constraints is implemented in all our algorithms using the "lazy cut callback." functionality available in Gurobi. The Python implementations for all approaches are publicly available in our GitHub repository (Vaghfi Mohebbi et al. 2026).

As we are only interested in $\ell$-frequent cliques, we preprocess the graph by deleting every vertex $u$ with $|A_u| \leqslant \ell - 1$ and every edge $uv$ with $|A_u \cap A_v| \leqslant \ell - 1$. Henceforth, we assume without loss of generality that $|A_u| \geqslant \ell$ for each $u \in V$ and $|A_u \cap A_v| \geqslant \ell$ for each $uv \in E$.

The following column headings are used in the tables reporting numerical results. The column labeled "Size" reports the patient cohort size. In summary tables, the quantities reported under "Avg" and "Range" are the average and range of the results for each cohort size across the five samples tested, respectively. The average and range of the best objective

value found are reported under "Best Obj," and the wall-clock running time (rounded down to the nearest second) is reported under "Time." An entry of "TO" for running time indicates that the algorithm terminated by reaching the 2-h time limit. The column "Gap" reports the percentage optimality gap at termination; a value of zero indicates optimal resolution within solver tolerance. In LBBD methods, columns "#Opt Cuts" and "#Feas Cuts" report the number of optimality and feasibility cuts added, respectively, and column "#CBB Cuts" reports the number of cuts generated by the hybrid LBBD algorithm.

### 7.3. Comparing Combinatorial Algorithms

Table 1 reports the summary of results comparing the modified Bron–Kerbosch (BK) clique enumeration algorithm from (Vaghfi Mohebbi et al. 2025) without clique size upper-bound enforcement and the CBB algorithm (Algorithm 2 in Appendix B). The most significant differentiator between the algorithms is that the BK algorithm enumerates all $\ell$-frequent cliques while the CBB algorithm prunes the recursive calls using the $\gamma$-upper-bound. We report the average decrease in the number of recursive calls made under the column labeled "#RC decrease", as the reduction is directly attributable to the use of $\gamma$-upper-bound (2). Detailed results for each sample are reported in Table 7 in Appendix C.

Both algorithms successfully solve all tested instances up to 30,000 patients. For the BK algorithm, average running times increase from less than 1 second for 1,000 patients to approximately 73 seconds for the 30,000-patient instances. The CBB algorithm is significantly faster than the modified BK algorithm. Its average running time remains below 1 second for instances with up to 10,000 patients. For larger instances, the running time increases to 3 seconds at 15,000 patients, 8 seconds at 20,000 patients, 22 seconds at 25,000 patients, and drops back to only 2 seconds for the 30,000-patient instances, while the BK algorithm requires 73 seconds on average for that cohort size. The (average) maximum mortality rate ranges from 0.45 at 1000 patients to 0.85 at 30,000 patients. The full 223,291-patient instance exceeds the time limit for both algorithms. The number of recursive calls in the BK algorithm reaches over one million for the largest solved instances (up to approximately 1.39 million in the 30,000-patient cohort), while CBB requires far fewer. Across the cohort sizes the average reduction in recursive calls varies from 28 to 1,153,091.

### 7.4. Comparing Direct MILP Solution

We implemented the MILP formulation (13) included in Appendix A using delayed addition of the linearizing constraints (13b)–(13d) that scale with $|P|$. We add all violated

**Table 1 Comparing the Modified Bron–Kerbosch and CBB Algorithms**

| Size | Best Obj | | BK, Time | | CBB, Time | | #RC decrease |
|---|---|---|---|---|---|---|---|
| | Avg | Range | Avg | Range | Avg | Range | Avg |
| 1,000 | 0.45 | [0.41, 0.49] | $<1$ | [$<1$, $<1$] | $<1$ | [$<1$, $<1$] | 28 |
| 5,000 | 0.80 | [0.72, 0.88] | 1 | [1, 1] | $<1$ | [$<1$, $<1$] | 2,207 |
| 10,000 | 0.86 | [0.81, 0.92] | 2 | [2, 3] | 1 | [1, 1] | 20,787 |
| 15,000 | 0.95 | [0.89, 1.00] | 8 | [5, 9] | 3 | [2, 4] | 89,191 |
| 20,000 | 0.97 | [0.96, 0.99] | 24 | [22, 27] | 8 | [7, 8] | 241,932 |
| 25,000 | 0.97 | [0.97, 0.98] | 47 | [42, 53] | 22 | [20, 26] | 478,683 |
| 30,000 | 0.85 | [0.83, 0.88] | 73 | [66, 83] | 2 | [2, 2] | 1,153,091 |
| 223,291 | UNK | UNK | TO | UNK | UNK | TO | UNK |

UNK stands for Unknown (no value was reported because the run timed out).

constraints of this type to BC nodes at which an integral solution is encountered as the node LP relaxation optimum. Preliminary experiments showed this choice to yield better performance over delayed addition of constraints (1b)–(1c). However, we explored an implementation that used an "aggregated" variant of constraints (1b), given by constraints (11), in the root LP formulation. Note that constraints (11) model the requirement exactly.

$$|V\backslash D_i|y_i \leqslant |V\backslash D_i| - \sum_{j\in V\backslash D_i} x_j \qquad \forall i \in P \tag{11}$$

We implemented formulation (14) based on the Charnes and Cooper (1962) transformation included in Appendix A using Gurobi's logical constraint feature for constraints (14b)–(14d), which essentially linearize the products $\bar{x}_u = x_u \cdot t$. Gurobi's indicator constraints can directly model the implication that, if $x_u = 1$ then $\bar{x}_u = t$, else $\bar{x}_u = 0$, avoiding the need for explicit big-M values (Gurobi Optimization, LLC 2025). We also add constraints (14e)–(14f) in a delayed manner; all violated constraints of this type are added at the BC node if an integral solution is encountered as the LP relaxation optimum.

Table 8 in Appendix C reports detailed results comparing our implementations of formulation (13) with aggregation, formulation (13) without aggregation, and formulation (14). All three formulations can solve the 1,000-patient instances to optimality. Across the samples, solving formulation (13) with aggregation takes 16–35 seconds, and 66–75 seconds without aggregation. By contrast, solving formulation (14) takes 2–4 seconds across the samples. Only the Charnes–Cooper transformation based formulation (14) consistently solves the 5,000-patient instances to optimality (0% gap, 863–3,770 seconds). Both implementation variants of formulation (13) reach the time limit with optimality gaps of 13%–54% with aggregation and 13%–47% without aggregation. For cohorts of 10,000 patients and larger, all three formulations reach the time limit on nearly all instances.

### 7.5. Comparing LBBD Variants

Recall our two LBBD variations, that we will refer to henceforth as "pure LBBD" and "hybrid LBBD" for convenience. Pure LBBD is the DBC algorithm described in Section 6.1 with its separation procedure described in Algorithm 3 in Appendix B invoked at BC nodes where the LP relaxation optimum is integral. Hybrid LBBD is described in Section 6.2, which uses the separation procedure described in Algorithm 1 instead, adding CBB cuts if the $\ell$-frequent clique encountered at the BC node contains at least $\tau$ vertices. Another notable difference is that the root relaxation is strengthened with CBB cuts based on optimal cliques found from $k$ calls to the CBB algorithm. Each call to the CBB algorithm is made with a disease among those with the $k$ highest individual mortality rate. Constraint (4f) is also improved, if possible, based on the best mortality rate among the cliques found. After preliminary experimentation, we set $\tau = 4$ and $k = 5$.

**Table 2 Summary of Results from the Pure LBBD Algorithm**

| Size | Gap | | Time | | #Opt Cuts | | #Feas Cuts | |
|---|---|---|---|---|---|---|---|---|
| | Avg | Range | Avg | Range | Avg | Range | Avg | Range |
| 1,000 | 0 | [0, 0] | $<1$ | [$<1$, $<1$] | 13.6 | [4, 21] | 52.4 | [18, 92] |
| 5,000 | 0 | [0, 0] | 3 | [2, 4] | 1,051 | [783, 1,365] | 778.4 | [577, 1,039] |
| 10,000 | 0 | [0, 0] | 296 | [240, 378] | 15,633 | [11,427, 19,593] | 14,211 | [12,036, 16,057] |
| 15,000 | 3 | [0, 6] | TO | [47, TO] | 60,683 | [5,010, 92,190] | 52,675 | [4,452, 74,083] |
| 20,000 | 3 | [1, 4] | TO | [TO, TO] | 72,174 | [62,069, 85,761] | 72,691 | [62,164, 82,328] |
| 25,000 | 3 | [2, 4] | TO | [TO, TO] | 65,231 | [49,369, 82,039] | 76,147 | [59,438, 89,502] |
| 30,000 | 3 | [3, 5] | TO | [TO, TO] | 61,358 | [56,441, 75,841] | 137,369 | [114,797, 150,718] |
| 223,291 | 3 | | TO | | 58,825 | | 42,595 | |

**Table 3 Summary of Results from the Hybrid LBBD Algorithm**

| Size | Gap | | Time | | #Opt Cuts | | #Feas Cuts | | #CBB Cuts | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Avg | Range | Avg | Range | Avg | Range | Avg | Range | Avg | Range |
| 1,000 | 0 | [0, 0] | $<1$ | [$<1$, $<1$] | 8.6 | [5, 14] | 0 | [0, 0] | 0 | [0, 0] |
| 5,000 | 0 | [0, 0] | 1 | [1, 2] | 912 | [736, 1,249] | 566 | [507, 645] | 63 | [4, 183] |
| 10,000 | 0 | [0, 0] | 100 | [64, 146] | 9,053 | [7,762, 10,571] | 7,062 | [4,764, 9,619] | 4,599 | [2,527, 6,842] |
| 15,000 | 0 | [0, 0] | 1,283 | [7, 2,394] | 26,169 | [577, 39,564] | 4,118 | [45, 7,257] | 21,731 | [189, 34,797] |
| 20,000 | 2 | [0, 4] | TO | [TO, TO] | 81,976 | [79,347, 89,610] | 72,336 | [70,127, 76,794] | 44,145 | [41,677, 48,648] |
| 25,000 | 2 | [1, 3] | TO | [TO, TO] | 74,992 | [62,595, 82,335] | 76,783 | [66,641, 86,002] | 43,782 | [37,470, 48,200] |
| 30,000 | 0 | [0, 0] | 716 | [532, 1,102] | 19,945 | [18,620, 22,737] | 24,769 | [23,116, 26,965] | 7,158 | [6,476, 8,336] |
| 223,291 | 2 | | TO | | 423 | | 1 | | 201 | |

Tables 2 and 3 present the summary of the results for the LBBD variants, and detailed sample-wise results are presented in Tables 9 and 10 in Appendix C. Both variants solve the 1,000-patient and 5,000-patient instances very quickly. On 10,000-patient instances, both

variants solve all five instances to optimality, although the hybrid LBBD algorithm takes a third of the time required by the pure LBBD algorithm, on average. For 15,000-patient instances, pure LBBD solves two instances to optimality (in 47 and 2,278 seconds) and times-out on the remaining three with optimality gaps of 4–6%. hybrid LBBD solves all five instances to optimality. Both variants time-out on all instances with small optimality gaps for 20,000 and 25,000-patient instances, although the hybrid LBBD algorithm reports slightly smaller gaps, on average. Notably, for 30,000-patient instances, while pure LBBD algorithm times out on all instances, the hybrid LBBD algorithm solves all of them to optimality. On the full 223,291-patient instance, both pure LBBD and hybrid LBBD algorithms time out, with the hybrid variant reporting a smaller optimality gap. Overall, the hybrid LBBD algorithm is substantially faster on instances up to 15,000 patients and is the only method that solves all 30,000-patient instances to optimality. Detailed numerical results for all instances and all algorithms are provided in Appendix C. A full list of disease abbreviations is provided in Appendix D.

## 8. Side Constraints and the Hybrid LBBD Algorithm: A Case Study

The effectiveness of the LBBD variants and the CBB algorithm suggests the upper-bound (2) has been helpful in various guises in these methods. The LBBD approaches however still take considerably longer time to solve the core optimization problem compared to the CBB algorithm, although it is vastly superior to using the two MILP formulations.

An important feature of the LBBD approaches is that it is based on integer programming, and therefore benefits from the expressiveness of the framework. Capturing additional constraints, possibly modeling context specific practical requirements, that can be expressed using the $x$ variables can be accomplished with relative ease by modifying the underlying formulation of $\mathcal{X}$. Although the CBB algorithm may require extensive modifications to solve the new problem, the CBB cut will still be valid as the (unmodified) CBB algorithm still computes a valid upper-bound on the optimal solution in the presence of new constraints in $\mathcal{X}$. The hybrid LBBD algorithm will continue to remain valid under these modifications. To demonstrate this point, we consider a case study with a side-constraint to the original problem based on diseases that predominantly afflict older patients.

Older patients typically exhibit more comorbidities and higher baseline mortality. As a result, the maximum mortality rate cliques may be heavily influenced by older patient

populations. So, we choose a cutoff of age 60, and for each disease compute the proportion of its patients who are 60 or older: $a_{60}(v) := |\{i \in A_v : \text{age}(i) \geqslant 60\}|/|A_v|, \forall v \in V$. We limit the average $a_{60}(v)$ across the selected diseases to be at most 0.5 in constraint (12). Our aim is to identify a maximum mortality rate clique after limiting the extent to which age composition can explain the observed lethality.

$$\sum_{v \in V} a_{60}(v) x_v \leqslant 0.5 \sum_{v \in V} x_v \tag{12}$$

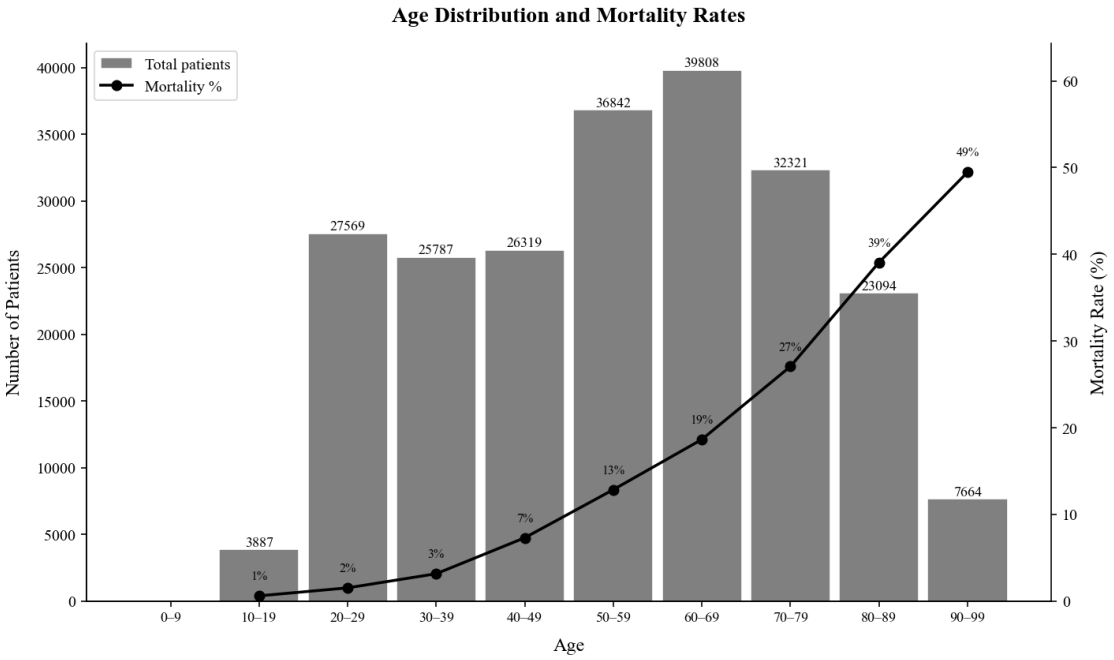


**Figure 1** **Age Distribution and Mortality Rates Among Patients in Our Dataset**

In our dataset, ages range from 18 to 91, with an average age of 56. Approximately 54% of patients are younger than 60. We note that patient ages in MIMIC-IV are derived from the `anchor_age` column in `patients.csv`, which caps all ages above 89 at 91 (Johnson et al. 2023). As our age threshold of 60 falls well below this ceiling, this grouping does not affect our analysis. The age distribution and mortality rates corresponding to each age group is presented in Figure 1, where the bins used are $[0, 9], \ldots, [90, 99]$.

We solve problem (3) with constraint (12) included in the description of $\mathcal{X}$ using the hybrid LBBD algorithm. We also exclude constraint (4f), which is no longer valid in the presence of constraint (12), rather than modify $\mu_{LB}$ to match. Our intention was to minimize tailoring of the methodology in the presence of the new constraint, as it is only meant

to be illustrative. Table 11 in Appendix C presents the detailed results of this study, across all instance sizes, which is summarized in Table 4. All instances across every size group are solved to optimality (0% Gap). The additional constraint appears to have made the instances more amenable to optimal resolution, compared against the results in Table 3.

**Table 4 Summary of Results from the Hybrid LBBD Algorithm with Age-Based Side-Constraint**

| Size | Time | | #Opt Cuts | | #Feas Cuts | | #CBB Cuts | |
|---|---|---|---|---|---|---|---|---|
| | Avg | Range | Avg | Range | Avg | Range | Avg | Range |
| 1,000 | $<1$ | [$<1$, $<1$] | 5 | [3, 7] | 0 | [0, 0] | 0 | [0, 0] |
| 5,000 | 0 | [0, 1] | 55 | [35, 78] | 34 | [14, 56] | 0 | [0, 0] |
| 10,000 | 5 | [3, 9] | 228 | [152, 273] | 189 | [94, 277] | 7 | [3, 13] |
| 15,000 | 36 | [12, 45] | 628 | [359, 742] | 595 | [320, 769] | 81 | [27, 114] |
| 20,000 | 113 | [89, 131] | 1,466 | [1,106, 1,788] | 1,472 | [948, 1,921] | 290 | [224, 350] |
| 25,000 | 290 | [226, 397] | 2,724 | [2,252, 3,537] | 2,836 | [2,101, 3,848] | 706 | [511, 992] |
| 30,000 | 23 | [18, 29] | 584 | [505, 766] | 1,015 | [889, 1,322] | 20 | [10, 31] |

Instances with 15,000 patients or less are solved in under a minute, with the smallest instances (1,000 and 5,000 patients) solved in under 1 second, requiring only optimality cuts and feasibility cuts, but no CBB cuts. For 20,000-patient and larger instances, we see many more CBB cuts being used and the running times remain reasonable. For the largest group solved with 30,000 patients, solution times are substantially shorter, ranging from 18 to 29 seconds. The hybrid LBBD appears to be markedly effective on the 30,000-patient instances based on our results with and without the side-constraint.

To see the impact of the side-constraint on the diseases detected, we present the cliques obtained by enforcing and not enforcing constraint (12) when solving samples of size 25,000 and 30,000 in Tables 5 and 6, respectively. The tables report results column-wise for each of the five samples. Under each sample column, the column "W/O" reports the solution found without constraint (12) and column "W" reports the solution found with constraint (12). In addition to reporting the clique of diseases found, we report the $a_{60}()$ value for each disease, as well as the mortality rate of the clique and the average over the $a_{60}()$ values in the two bottom rows. Tables 12 and 13 in Appendix C present a summary, average and range of the $a_{60}$ values for all diseases appearing in Tables 5 and 6 across different samples of sizes 25,000 and 30,000, respectively.

The solutions obtained without the side-constraint exhibit a consistent pattern across all instances: every disease appearing in these cliques has $a_{60}(v) \geqslant 0.5$, meaning the majority of its patient population is aged 60 or older. For the 25,000-sample instances, the

**Table 5 Optimal Solutions Found With and Without the Age-Based Side-Constraint on 25,000-Patient Samples**

| Sample #1 | | | | Sample #2 | | | | Sample #3 | | | | Sample #4 | | | | Sample #5 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| W/O | | W | | W/O | | W | | W/O | | W | | W/O | | W | | W/O | | W | |
| Clique | $a_{60}$ | Clique | $a_{60}$ | Clique | $a_{60}$ | Clique | $a_{60}$ | Clique | $a_{60}$ | Clique | $a_{60}$ | Clique | $a_{60}$ | Clique | $a_{60}$ | Clique | $a_{60}$ | Clique | $a_{60}$ |
| OACE | 0.72 | RCU | 0.56 | CA | 0.79 | ONSD | 0.56 | NCD | 0.88 | PIA | 0.48 | OACE | 0.71 | Hepatitis | 0.28 | SEPT | 0.61 | DCP | 0.58 |
| RF | 0.66 | FevUO | 0.44 | OACE | 0.72 | IntI | 0.54 | OACE | 0.72 | OLD | 0.47 | HF | 0.79 | OLD | 0.47 | OACE | 0.72 | AlcD | 0.22 |
| PNA | 0.63 | OLD | 0.48 | DLM | 0.73 | MD | 0.38 | AURF | 0.68 | | | PDX-Unacc | 0.50 | ARenF | 0.71 | DLM | 0.73 | OLD | 0.48 |
| OSNSD | 0.68 | | | RF | 0.66 | | | PDX-Unacc | 0.51 | | | SHK | 0.68 | | | PDX-Unacc | 0.50 | ARenF | 0.71 |
| PDX-Unacc | 0.51 | | | HF | 0.79 | | | OSNSD | 0.67 | | | FED | 0.61 | | | CoagHD | 0.60 | | |
| | | | | PDX-Unacc | 0.51 | | | | | | | | | | | | | | |
| $\mu$ | Avg | $\mu$ | Avg | $\mu$ | Avg | $\mu$ | Avg | $\mu$ | Avg | $\mu$ | Avg | $\mu$ | Avg | $\mu$ | Avg | $\mu$ | Avg | $\mu$ | Avg |
| 0.97 | 0.64 | 0.68 | 0.49 | 0.97 | 0.70 | 0.67 | 0.49 | 0.98 | 0.69 | 0.69 | 0.48 | 0.97 | 0.66 | 0.68 | 0.49 | 0.98 | 0.63 | 0.70 | 0.50 |

Disease names are abbreviated due to space limitations; full names are provided in Appendix D.

**Table 6 Optimal Solutions Found With and Without the Age-Based Side-Constraint on 30,000-Patient Samples**

| Sample #1 | | | | Sample #2 | | | | Sample #3 | | | | Sample #4 | | | | Sample #5 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| W/O | | W | | W/O | | W | | W/O | | W | | W/O | | W | | W/O | | W | |
| Clique | $a_{60}$ | Clique | $a_{60}$ | Clique | $a_{60}$ | Clique | $a_{60}$ | Clique | $a_{60}$ | Clique | $a_{60}$ | Clique | $a_{60}$ | Clique | $a_{60}$ | Clique | $a_{60}$ | Clique | $a_{60}$ |
| HCSH | 0.79 | RCU | 0.55 | CHF-NH | 0.82 | AlcD | 0.20 | DOA | 0.59 | PS | 0.49 | CardArr | 0.79 | AlcD | 0.21 | HCSH | 0.80 | AlcD | 0.21 |
| ARF | 0.62 | DOA | 0.59 | DOA | 0.57 | OLD | 0.50 | CardArr | 0.79 | | | HCSH | 0.79 | OLD | 0.51 | SHK | 0.68 | OLD | 0.50 |
| ONSD | 0.56 | AlcD | 0.22 | CUS | 0.74 | | | GD | 0.59 | | | CHF-NH | 0.82 | | | OA | 0.62 | | |
| RCU | 0.55 | | | | | | | RCU | 0.57 | | | CUS | 0.73 | | | | | | |
| | | | | | | | | CA | 0.81 | | | | | | | | | | |
| $\mu$ | Avg | $\mu$ | Avg | $\mu$ | Avg | $\mu$ | Avg | $\mu$ | Avg | $\mu$ | Avg | $\mu$ | Avg | $\mu$ | Avg | $\mu$ | Avg | $\mu$ | Avg |
| 0.83 | 0.63 | 0.60 | 0.45 | 0.88 | 0.71 | 0.65 | 0.35 | 0.87 | 0.67 | 0.61 | 0.49 | 0.84 | 0.78 | 0.66 | 0.36 | 0.85 | 0.70 | 0.64 | 0.36 |

Disease names are abbreviated due to space limitations; full names are provided in Appendix D.

diseases driving the optimal cliques include neurocognitive disorders ($a_{60} = 0.88$), congestive heart failure ($a_{60} = 0.79$), coronary atherosclerosis ($a_{60} = 0.79$), disorders of lipid metabolism ($a_{60} = 0.73$), and other aftercare encounters ($a_{60} = 0.72$). The 30,000-sample instances exhibit a similar pattern: congestive heart failure ($a_{60} = 0.82$), hypertension with complications ($a_{60} = 0.80$), coronary atherosclerosis ($a_{60} = 0.81$), and cardiac arrhythmia ($a_{60} = 0.79$) dominate. The optimal cliques identified without the side-constraint achieve mortality rates in the range $[0.97, 0.98]$ for the 25,000-patient instances and in the range $[0.83, 0.88]$ for the 30,000-patient instances, and are heavily populated by older skewing diseases as the average $a_{60}()$ value is more than 0.6 in all instances. These results suggest that, without the side-constraint limiting the proportion of older skewing diseases, the diseases in the maximum mortality rate cliques may conflate two distinct phenomena: the elevated baseline mortality of older patients and the potentially synergistic lethality of co-occurring diseases. In effect, these solutions are partly driven by diseases that cluster

in older patients rather than exclusively reflecting disease combinations that are intrinsically dangerous independent of age effects. By contrast, optimal solutions obtained with the side-constraint incorporate diseases with $a_{60} < 0.5$, including alcohol-related disorders ($a_{60} = 0.22$), hepatitis ($a_{60} = 0.28$), mood disorders ($a_{60} = 0.38$), and fever of unknown origin ($a_{60} = 0.44$). Even after limiting the proportion of older-skewing diseases, mortality rate remains relatively high. Specifically, the mortality rates are in the range $[0.67, 0.70]$ for the 25,000-patient instances and in the range $[0.60, 0.66]$ for the 30,000-patient instances, indicating that severe outcomes are not contingent on cardiovascular or respiratory failures but can arise from non-cardiac disease interactions.

In the presence of age-based side constraints, we observe that high-mortality cliques found frequently include alcohol-related disorders, other liver diseases, and acute renal failure. While our study does not attempt to infer clinical mechanisms, this observation aligns with patterns previously reported in clinical studies, which have noted that alcohol-related liver disease co-occurring with acute kidney injury is often associated with elevated short-term mortality across adult age ranges, rather than being exclusively concentrated in elderly populations (Altamirano et al. 2012).

## 9. Conclusion

This article establishes the NP-hardness and introduces new exact methods for solving the maximum mortality rate clique problem using integer programming and combinatorial optimization techniques, culminating in a tailored hybrid logic-based Benders decomposition algorithm. This optimization framework enables the identification of lethal disease clusters from EHR data, offering practical advantages and optimality guarantees.

Our computational study, conducted on 35 sampled instances with cohort sizes ranging from 1,000 to 30,000 patients, as well as the full 223,291-patient publicly available MIMIC-IV EHR dataset, demonstrates the following. The combinatorial branch-and-bound algorithm substantially outperforms the modified Bron–Kerbosch algorithm introduced in (Vaghfi Mohebbi et al. 2025), reducing recursive calls by over a million on the largest instances and solving all tested cohort sizes to optimality. Both LBBD variants are competitive, while the hybrid variant provides significantly faster solution times on smaller and medium-sized instances and smaller optimality gaps on the larger instances.

We further demonstrate the flexibility of the hybrid variant to incorporate side-constraints on the cliques. As an example, we introduce an age-based constraint that

limits the proportion of patients aged 60 or older across the diseases in the selected clique. Incorporating this constraint into the model reveals a qualitatively different class of lethal cliques centered on systemic and metabolic disease interactions, such as combinations of alcohol-related disorders, liver disease, and acute renal failure, that are unlikely to be explained primarily by the elevated baseline mortality of elderly patients. Importantly, such side-constraints can be seamlessly incorporated into the hybrid LBBD framework without modifying the combinatorial branch-and-bound algorithm, underscoring its advantage.

We do recognize some limitations of our study. The MIMIC-IV dataset is derived from a single academic medical center, which may limit the generalizability of the specific disease clusters identified. We note that while the proposed methods were initially developed and evaluated on another EHR dataset comprising millions of patient encounters, access to this dataset was withdrawn during the development of this article; consequently, those results are not included in the present study. The one-year post-discharge mortality definition used in MIMIC-IV means that findings may differ under alternative mortality windows or definitions, such as in-hospital mortality alone. As with any study based on coded diagnoses, the quality of results is sensitive to the completeness and accuracy of ICD coding practices. External validation using diverse EHR systems and healthcare settings would strengthen confidence in the robustness of our methods and the results.

In the future, it may be worth extending the problem setting to explicitly incorporate the temporal information on patient encounters through temporal comorbidity graphs (Lu et al. 2021). The notion of introducing side-constraints in the disease-variable space can be extended to encode other clinically important requirements based on demographic information enabling targeted investigations of specific patient subpopulations. Finally, applying the framework to larger, multi-institution EHR datasets would support broader validation and discovery of previously understudied lethal disease combinations.

## Acknowledgments

The computing for this project was performed at the High Performance Computing Center at Oklahoma State University (HPC Pete) supported in part through the National Science Foundation grant OAC-1126330.

## References

Agency for Healthcare Research and Quality (2021) Software. Healthcare Cost and Utilization Project (HCUP).

Altamirano J, Fagundes C, Dominguez M, García E, Michelena J, Cárdenas A, Guevara M, Pereira G, Torres-Vigil K, Arroyo V, et al. (2012) Acute kidney injury is an early predictor of mortality for patients with alcoholic hepatitis. *Clinical Gastroenterology and Hepatology* 10(1):65–71.

Balasubramanian H, Prakash S, Jafari A, Mohan A, Gopalappa C (2024) *Interventions for Patients with Complex Medical and Social Needs*, chapter 12, 358–391 (INFORMS).

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

Cappanera P, Gavanelli M, Nonato M, Roma M (2023) Logic-based Benders decomposition in answer set programming for chronic outpatients scheduling. *Theory and Practice of Logic Programming* 23(4):848–864.

Centers for Disease Control and Prevention (2013) International classification of diseases, 9th revision, clinical modification (ICD-9-CM). URL: https://www.cdc.gov/nchs/icd/icd9cm.htm, Accessed: April 2026.

Centers for Disease Control and Prevention (2022) International classification of diseases, 10th revision, clinical modification (ICD-10-CM). URL: https://www.cdc.gov/nchs/icd/icd-10-cm.htm, Accessed: April 2026.

Charnes A, Cooper WW (1962) Programming with linear fractional functionals. *Naval Research Logistics Quarterly* 9(3-4):181–186.

Codato G, Fischetti M (2006) Combinatorial Benders' cuts for mixed-integer linear programming. *Operations Research* 54(4):756–766.

Côté JF, Haouari M, Iori M (2021) Combinatorial Benders decomposition for the two-dimensional bin packing problem. *INFORMS Journal on Computing* 33(3):963–978.

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

Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE (2000) Physiobank, PhysioToolkit, and PhysioNet. *Circulation* 101(23):e215–e220.

Guo C, Bodur M, Aleman DM, Urbach DR (2021) Logic-based Benders decomposition and binary decision diagram based approaches for stochastic distributed operating room scheduling. *INFORMS Journal on Computing* 33(4):1551–1569.

Gurobi Optimization, LLC (2025) How do I model general conditional statements in gurobi? URL: https://support.gurobi.com/hc/en-us/articles/4414392016529-How-do-I-model-general-conditional-statements-in-Gurobi, Accessed: April 2026.

Hooker J (2023) *Logic-based Benders decomposition: theory and applications* (Springer Nature), ISBN 978-3-031-45039-6.

Jensen PB, Jensen LJ, Brunak S (2012) Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews Genetics* 13(6):395–405.

Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, Pollard TJ, Hao S, Moody B, Gow B, Lehman LwH, Celi LA, Mark RG (2023) MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data* 10(1):1, ISSN 2052-4463.

Juvin C, Houssin L, Lopez P (2023) Logic-based Benders decomposition for the preemptive flexible job-shop scheduling problem. *Computers & Operations Research* 152:106156.

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.

Karlsson E, Rönnberg E (2022) Logic-based Benders decomposition with a partial assignment acceleration technique for avionics scheduling. *Computers & Operations Research* 146:105916.

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.

Malecki SL, Loffler A, Tamming D, Johansen ND, Biering-Sørensen T, Fralick M, Sohail S, Shi J, Roberts SB, Colacci M, et al. (2024) Development and external validation of tools for categorizing diagnosis codes in international hospital data. *International Journal of Medical Informatics* 105508.

Michels AS, Lopes TC, Sikora CGS, Magatao L (2019) A Benders' decomposition algorithm with combinatorial cuts for the multi-manned assembly line balancing problem. *European Journal of Operational Research* 278(3):796–808.

Mohebbi PV, Salehiyan A, Deep A (2025) Disease cluster analysis in electronic health records: Insights into mortality and comorbidity patterns. *IISE Annual Conference Proceedings* 1–6, DOI: 10.21872/2025IISE_6965.

Ochiai A (1957) Zoogeographic studies on the Soleoid fishes found in Japan and its neighbouring regions. *Bulletin of Japanese Society of Scientific Fisheries* 22:526–530.

Redelmeier DA, Tan SH, Booth GL (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 MJ (1983) *Introduction to Modern Information Retrieval* (New York, NY, USA: McGraw-Hill, Inc.).

Sarwar T, Seifollahi S, Chan J, Zhang X, Aksakalli V, Hudson I, Verspoor K, Cavedon L (2022) The secondary use of electronic health records for data mining: Data characteristics and challenges. *ACM Computing Surveys* 55(2).

Shickel B, Tighe PJ, Bihorac A, Rashidi P (2018) Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE Journal of Biomedical and Health Informatics* 22(5):1589–1604.

Vaghfi Mohebbi P, Lu Y, Balasundaram B (2026) Implementation codes of Charnes–Cooper and LBBD Methods. URL: https://gitfront.io/r/Parisam/HyoiYiV5kZoe/IJOC-MRC/. Accessed: April 2026.

Vaghfi Mohebbi P, Lu Y, Miao Z, Balasundaram B, Kalgotra P, Sharda R (2025) Identifying most lethal cliques in disease comorbidity graphs. *IISE Transactions on Healthcare Systems Engineering* 15(2):183–200.

Yadav P, Steinbach M, Kumar V, Simon G (2018) Mining electronic health records (EHRs): A survey. *ACM Computing Surveys* 50(6).

Zhong H, Loukides G, Pissis SP (2022) Clustering demographics and sequences of diagnosis codes. *IEEE Journal of Biomedical and Health Informatics* 26(5):2351–2359.

## Appendix A: MILP Formulations

Vaghfi Mohebbi et al. (2025) introduced a continuous variable $z$ to model the fractional objective and auxiliary continuous variables $w_i \in [0,1]$ to represent the products $zy_i$. This formulation is presented next.

$$\max\ z \tag{13a}$$

$$s.t.\quad w_i \leqslant y_i \qquad \forall i \in P \tag{13b}$$

$$w_i \geqslant z + y_i - 1 \qquad \forall i \in P \tag{13c}$$

$$w_i \leqslant z \qquad \forall i \in P \tag{13d}$$

$$\sum_{i \in \tilde{P}} y_i = \sum_{i \in P} w_i \tag{13e}$$

$$z \in [0,1] \tag{13f}$$

$$w_i \in [0,1] \qquad \forall i \in P \tag{13g}$$

$$(1b) - (1e)$$

Constraints (13b)–(13d) force $w_i = zy_i$, then equation (13e) correctly captures the objective function value in $z$.

An alternative way to linearize fractional programs with nonnegative variables and a positive denominator is the Charnes and Cooper (1962) transformation, which we employ in the following reformulation.

$$\max\ \sum_{i \in \tilde{P}} \bar{y}_i \tag{14a}$$

$$s.t.\quad \bar{x}_u \geqslant t - \frac{1}{|P|}(1 - x_u) \qquad \forall u \in V \tag{14b}$$

$$\bar{x}_u \leqslant \frac{1}{\ell} x_u \qquad \forall u \in V \tag{14c}$$

$$\bar{x}_u \leqslant t \qquad \forall u \in V \tag{14d}$$

$$\bar{y}_i \leqslant t - \bar{x}_u \qquad \forall u \in V \backslash D_i,\ i \in P \tag{14e}$$

$$\bar{y}_i \geqslant t - \sum_{u \in V \backslash D_i} \bar{x}_u \qquad \forall i \in P \tag{14f}$$

$$\sum_{i \in P} \bar{y}_i = 1 \tag{14g}$$

$$\bar{x}_u \geqslant 0 \qquad \forall u \in V \tag{14h}$$

$$0 \leqslant \bar{y}_i \leqslant \frac{1}{\ell} \qquad \forall i \in P \tag{14i}$$

$$\frac{1}{|P|} \leqslant t \leqslant \frac{1}{\ell} \tag{14j}$$

$$x \in \mathcal{X} \tag{14k}$$

This reformulation introduces a continuous variable $t$ that is intended to model $t \sum_{i \in P} y_i = 1$, representing the reciprocal of the denominator of the mortality rate function. As we are only interested in $\ell$-frequent cliques, constraint (14j) places the trivial bounds on $t$. We retain the original $x$ variables as the clique incidence vectors with $x \in \mathcal{X}$ and introduce continuous variables $\bar{x}_u$ for $u \in V$ to model $\bar{x}_u = tx_u$ using constraints (14b)–(14d). If $x_u = 1$, constraints (14b) and (14c) force $\bar{x}_u = t$ while constraint (14c) is redundant

in the presence of the upper bound in constraint (14j). If $x_u = 0$, then constraints (14c) and (14h) force $\bar{x}_u = 0$. In this case, constraints (14b) and (14d) are redundant in presence of the lower bound in constraint (14j). With $\bar{x}$ correctly modeled, we can now verify that $\bar{y}_i = ty_i$ for $i \in P$ is correctly modeled by constraints (14e) and (14f). If $x_u = 1$ for some $u \in V \backslash D_i$ and $i \in P$, then $\bar{x}_u = t$, and constraint (14e) forces $\bar{y}_i = 0$, while constraint (14f) is redundant as $\bar{y}_i$ is non-negative. However, if for some $i \in P$, $x_u = \bar{x}_u = 0$ for all $u \in V \backslash D_i$, then constraints (14e) and (14f) force $\bar{y}_i = t$. Now, we can conclude that in the presence of the scaling constraint (14g), the objective function (14a) correctly maximizes the mortality rate.

## Appendix B: Pseudocodes

**Algorithm 2:** Combinatorial Branch-and-Bound Algorithm for the Maximum Mortality Rate Clique Problem

**Input:** $\langle P, \widetilde{P}, A, G = (V, E), \ell \rangle$ and $\ell$-frequent clique $C^0$ (possibly empty)

**Output:** An $\ell$-frequent clique containing $C^0$ with maximum mortality rate

Initialize global variables $\texttt{BestSol} \leftarrow C^0$, $\texttt{BestObj} \leftarrow \mu(C^0)$
**if** $C^0 = \varnothing$ **then**
  $v \leftarrow \arg\max_{u \in V} \mu(u)$
  $\texttt{BestSol} \leftarrow \{v\}, \texttt{BestObj} \leftarrow \mu(v)$
  $\texttt{DfsBB}(\varnothing, V, P, \widetilde{P})$
**else**
  $L \leftarrow \bigcap_{u \in C^0} N(u) \backslash C^0$
  $\texttt{DfsBB}(C^0, L, \mathrm{den}(C^0), \mathrm{num}(C^0))$
**return** BestSol and BestObj
**Function** $\texttt{DfsBB}(C, L, \mathit{CurrentDen}, \mathit{CurrentNum})$
  **if** $L = \varnothing$ **then**
    **return**
  **for** $v \in L$ **do**
    $\texttt{NewDen} \leftarrow \texttt{CurrentDen} \cap A_v$
    $\texttt{NewNum} \leftarrow \texttt{CurrentNum} \cap A_v$
    **if** $|\mathit{NewDen}| \geqslant \ell$ *and* $\frac{|\mathit{NewNum}|}{\ell} > \mathit{BestObj}$ **then**
      **if** $\frac{|\mathit{NewNum}|}{|\mathit{NewDen}|} > \mathit{BestObj}$ **then**
        $\texttt{BestObj} \leftarrow \frac{|\texttt{NewNum}|}{|\texttt{NewDen}|}$
        $\texttt{BestSol} \leftarrow C \cup \{v\}$
      $\texttt{DfsBB}(C \cup \{v\}, L \cap N(v), \mathit{NewDen}, \mathit{NewNum})$
    $L \leftarrow L \backslash \{v\}$
  **return**

**Algorithm 3:** Logic-Based Benders Decomposition: Separation Procedure

**Input:** $\widetilde{x} \in \mathcal{X}$ and $\widetilde{z} \in [0,1]$

**Output:** Violated optimality or feasibility cut, if one exists

**if** $J_1(\widetilde{x})$ *is not* $\ell$*-frequent* **then**
  $\widetilde{C} \leftarrow$ `Minimalize`$(J_1(\widetilde{x}))$
  **return** violated feasibility cut: $\sum_{u \in \widetilde{C}} x_u \leqslant |\widetilde{C}| - 1$
**else if** $\widetilde{z} > \mu(\widetilde{x})$ **then**
  **return** violated optimality cut: $z \leqslant \mu(\widetilde{x}) + (1 - \mu(\widetilde{x})) \times \mathcal{H}(x, \widetilde{x})$
**else**
  **return** no violated cuts found

**Algorithm 4:** Clique Minimalization

**Input:** Clique $\widetilde{C}$ that is not $\ell$-frequent

**Output:** Minimal non-$\ell$-frequent clique

**Function** `Minimalize`$(\widetilde{C})$
  **repeat**
    `stopflag` $=$ `true`
    **for** $u \in \widetilde{C}$ **do**
      **if** $|\mathrm{den}(\widetilde{C} \backslash \{u\})| < \ell$ **then**
        $\widetilde{C} \leftarrow \widetilde{C} \backslash \{u\}$
        `stopflag` $=$ `false`
  **until** *stopflag* $=$ *true*;
  **return** $\widetilde{C}$

## Appendix C: Detailed Computational Results

The following column headings are used in the appendix tables reporting numerical results. "Size" denotes the patient cohort size, and "No." identifies the sample index. "Best Obj" reports the best objective value obtained, while "Time" gives the wall-clock running time, rounded down to the nearest second; "TO" indicates termination at the 2-hour time limit. "Gap" reports the percentage optimality gap at termination, where zero indicates that an optimal solution was obtained within solver tolerance. The number of branch-and-cut nodes explored is reported under "#BC Nodes" and column "#RC" indicates the number of recursive calls. Column "#CBB Calls" reports the number of calls to the CBB algorithm and "#CBB Rec" reports the total number of recursive calls. Finally, "#Opt Cuts" and "#Feas Cuts" denote the numbers of optimality and feasibility cuts added, respectively, while "#CBB Cuts" reports the number of CBB cuts generated by the hybrid LBBD algorithm. The "Clique" column lists the diseases forming the maximum clique identified for each sample. Column "Avg" reports the average value across the five samples and "Range" gives the minimum and maximum values observed.

**Table 7 Comparing the Modified Bron–Kerbosch Algorithm and CBB Algorithm**

| Size | No. | Clique | Best Obj | BK | | CBB | |
|---|---|---|---|---|---|---|---|
| | | | | Time | #RC | Time | #RC |
| 1,000 | 1 | PDX-Unacc, AURF | 0.45 | $<1$ | 40 | $<1$ | 14 |
| | 2 | ARF | 0.49 | $<1$ | 48 | $<1$ | 15 |
| | 3 | RCU, OA | 0.46 | $<1$ | 44 | $<1$ | 15 |
| | 4 | FED, PDX-Unacc | 0.41 | $<1$ | 43 | $<1$ | 20 |
| | 5 | FD | 0.43 | $<1$ | 40 | $<1$ | 10 |
| 5,000 | 1 | MALN, PDX-Unacc, OACE | 0.87 | 1 | 4,580 | $<1$ | 689 |
| | 2 | OACE, AURF, FED | 0.88 | 1 | 3,893 | $<1$ | 547 |
| | 3 | CHF-NH, ARF, CKD | 0.72 | 1 | 3,136 | $<1$ | 454 |
| | 4 | SM | 0.77 | 1 | 2,623 | $<1$ | 418 |
| | 5 | FD, CHF-NH, CardArr | 0.74 | 1 | 2,542 | $<1$ | 492 |
| 10,000 | 1 | PDX-Unacc, PNA, FED, OACE | 0.92 | 3 | 31,562 | 1 | 6,052 |
| | 2 | PDX-Unacc, DLM, CD, OACE | 0.92 | 3 | 31,545 | 1 | 5,324 |
| | 3 | SM, RCU, ONSD | 0.83 | 2 | 23,101 | 1 | 4,954 |
| | 4 | OA, SM | 0.83 | 2 | 21,784 | 1 | 4,730 |
| | 5 | SM, RCU, DOA | 0.81 | 2 | 20,643 | 1 | 3,640 |
| 15,000 | 1 | OACE, SHK | 0.95 | 9 | 122,916 | 4 | 21,387 |
| | 2 | SEPT, OSNSD, PDX-Unacc, RF, OACE | 0.94 | 9 | 126,861 | 4 | 21,775 |
| | 3 | OSNSD, AURF, PDX-Unacc, RF, OACE | 1.00 | 7 | 97,075 | 3 | 13,593 |
| | 4 | SEPT, CD, PDX-Unacc, FED, OACE | 0.96 | 8 | 117,050 | 4 | 19,757 |
| | 5 | DOA, ONSD, SM | 0.89 | 5 | 71,752 | 2 | 13,183 |
| 20,000 | 1 | OACE, SHK, RF | 0.97 | 27 | 332,591 | 8 | 61,953 |
| | 2 | OACE, AURF, COPD-BE, PDX-Unacc, FED | 0.96 | 26 | 312,853 | 8 | 57,524 |
| | 3 | OACE, RF, PNA, AURF, PDX-Unacc | 0.99 | 23 | 271,897 | 7 | 47,423 |
| | 4 | OACE, AURF, SHK, PDX-Unacc, FED | 0.96 | 25 | 301,779 | 8 | 58,434 |
| | 5 | OACE, DLM, RF, PDX-Unacc, CoagHD | 0.97 | 22 | 259,805 | 7 | 43,931 |
| 25,000 | 1 | OACE, RF, PNA, OSNSD, PDX-Unacc | 0.97 | 53 | 702,379 | 26 | 150,611 |
| | 2 | CA, OACE, DLM, RF, HF, PDX-Unacc | 0.97 | 47 | 615,018 | 23 | 134,531 |
| | 3 | NCD, OACE, AURF, PDX-Unacc, OSNSD | 0.98 | 42 | 541,644 | 20 | 112,198 |
| | 4 | OACE, HF, PDX-Unacc, SHK, FED | 0.97 | 46 | 595,396 | 23 | 132,610 |
| | 5 | SEPT, OACE, DLM, PDX-Unacc, CoagHD | 0.98 | 45 | 576,175 | 20 | 107,244 |
| 30,000 | 1 | HCSH, ARF, ONSD, RCU | 0.83 | 83 | 1,387,200 | 2 | 14,725 |
| | 2 | CHF-NH, DOA, CUS | 0.88 | 78 | 1,280,684 | 2 | 11,290 |
| | 3 | DOA, CardArr, GD, RCU, CA | 0.87 | 66 | 1,032,733 | 2 | 12,693 |
| | 4 | CardArr, HCSH, CHF-NH, CUS | 0.84 | 67 | 1,064,147 | 2 | 11,518 |
| | 5 | HCSH, SHK, OA | 0.85 | 70 | 1,061,875 | 2 | 10,960 |
| 223,291 | | - | - | TO | - | TO | - |

Disease names are abbreviated due to space constraints; full names are provided in Appendix D.

A dash "-" indicates that no value was reported by the solver for that metric.

**Table 8 Comparing Direct Solution of MILP Formulations**

| Size | No. | Formulation (13) - Aggregation | | | Formulation (13) - No Aggregation | | | Formulation (14) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Gap | Time | #BC Nodes | Gap | Time | #BC Nodes | Gap | Time | #BC Nodes |
| 1,000 | 1 | 0 | 17 | 1 | 0 | 71 | 3 | 0 | 2 | 215 |
| | 2 | 0 | 16 | 349 | 0 | 73 | 1 | 0 | 3 | 242 |
| | 3 | 0 | 34 | 505 | 0 | 68 | 1 | 0 | 4 | 415 |
| | 4 | 0 | 21 | 517 | 0 | 75 | 1 | 0 | 3 | 176 |
| | 5 | 0 | 35 | 1 | 0 | 66 | 1 | 0 | 4 | 198 |
| 5,000 | 1 | 14 | TO | 1,535,817 | 14 | TO | 6,256 | 0 | 1,033 | 10,118 |
| | 2 | 13 | TO | 764,563 | 13 | TO | 6,550 | 0 | 1,450 | 11,551 |
| | 3 | 54 | TO | 708,721 | 38 | TO | 5,941 | 0 | 863 | 5,624 |
| | 4 | 36 | TO | 2,916,160 | 45 | TO | 9,894 | 0 | 850 | 10,519 |
| | 5 | 41 | TO | 1,075,728 | 47 | TO | 5,440 | 0 | 3,770 | 9,505 |
| 10,000 | 1 | 8 | TO | 8,492 | - | TO | 0 | 8 | TO | 7,191 |
| | 2 | 9 | TO | 4,861 | 622 | TO | 1 | 8 | TO | 6,054 |
| | 3 | 38 | TO | 4,307 | 95 | TO | 1 | 24 | TO | 9,682 |
| | 4 | 20 | TO | 4,735 | 84 | TO | 1 | 20 | TO | 29,100 |
| | 5 | 67 | TO | 7,195 | 103 | TO | 1 | 0 | 6,986 | 24,184 |
| 15,000 | 1 | 57 | TO | 7,076 | - | TO | 0 | 5 | TO | 5,906 |
| | 2 | 23 | TO | 5,266 | - | TO | 0 | 19 | TO | 5,530 |
| | 3 | 6 | TO | 5,867 | 381 | TO | 1 | 0 | 3,554 | 2,342 |
| | 4 | 32 | TO | 6,516 | 162 | TO | 1 | 4 | TO | 5,171 |
| | 5 | 50 | TO | 6,431 | 576 | TO | 1 | 28 | TO | 5,795 |
| 20,000 | 1 | 23 | TO | 4,469 | - | TO | 0 | 31 | TO | 3,041 |
| | 2 | 51 | TO | 5,943 | - | TO | 0 | 20 | TO | 3,913 |
| | 3 | 11 | TO | 4,004 | - | TO | 0 | 14 | TO | 3,621 |
| | 4 | 4 | TO | 5,054 | - | TO | 0 | 7 | TO | 2,868 |
| | 5 | 33 | TO | 6,733 | 417 | TO | 1 | 71 | TO | 1,403 |
| 25,000 | 1 | 24 | TO | 7,350 | - | TO | 0 | 81 | TO | 2,758 |
| | 2 | 88 | TO | 50 | - | TO | 0 | 7 | TO | 1,319 |
| | 3 | 25 | TO | 40 | - | TO | 0 | 186 | TO | 877 |
| | 4 | 9 | TO | 5,643 | - | TO | 0 | 145 | TO | 1,712 |
| | 5 | 5 | TO | 6,411 | - | TO | 0 | 70 | TO | 1,824 |
| 30,000 | 1 | 51 | TO | 5,358 | - | TO | 0 | 87 | TO | 1,908 |
| | 2 | 38 | TO | 218 | - | TO | 0 | 254 | TO | 1,360 |
| | 3 | 476 | TO | 25 | - | TO | 0 | 154 | TO | 2,040 |
| | 4 | 50 | TO | 4,284 | - | TO | 0 | 37 | TO | 1,564 |
| | 5 | 23 | TO | 1,002 | - | TO | 0 | 32 | TO | 655 |
| 223,291 | | - | TO | 0 | - | TO | 0 | - | TO | 0 |

A dash "-" indicates that no value was reported by the solver for that metric.

**Table 9 Results of Pure LBBD Algorithm (Algorithm 3) for Different Instance Sizes**

| Size | No. | Algorithm 3 | | | |
|---|---|---|---|---|---|
| | | Gap | Time | #Opt Cuts | #Feas Cuts |
| 1,000 | 1 | 0 | $<1$ | 14 | 23 |
| | 2 | 0 | $<1$ | 17 | 50 |
| | 3 | 0 | $<1$ | 4 | 79 |
| | 4 | 0 | $<1$ | 21 | 92 |
| | 5 | 0 | $<1$ | 12 | 18 |
| 5,000 | 1 | 0 | 4 | 1,365 | 577 |
| | 2 | 0 | 3 | 963 | 639 |
| | 3 | 0 | 2 | 1,060 | 909 |
| | 4 | 0 | 3 | 1,085 | 1,039 |
| | 5 | 0 | 2 | 783 | 728 |
| 10,000 | 1 | 0 | 378 | 19,593 | 15,620 |
| | 2 | 0 | 347 | 19,204 | 16,057 |
| | 3 | 0 | 264 | 14,427 | 14,480 |
| | 4 | 0 | 240 | 13,515 | 12,863 |
| | 5 | 0 | 252 | 11,427 | 12,036 |
| 15,000 | 1 | *5 | TO | 77,900 | 67,228 |
| | 2 | *6 | TO | 79,463 | 70,431 |
| | 3 | 0 | 47 | 5,010 | 4,452 |
| | 4 | *4 | TO | 92,190 | 74,083 |
| | 5 | 0 | 2,278 | 48,852 | 47,180 |
| 20,000 | 1 | 4 | TO | 62,069 | 69,195 |
| | 2 | *4 | TO | 64,757 | 82,328 |
| | 3 | *1 | TO | 70,642 | 75,180 |
| | 4 | *4 | TO | 77,643 | 74,588 |
| | 5 | *3 | TO | 85,761 | 62,164 |
| 25,000 | 1 | 4 | TO | 64,476 | 69,818 |
| | 2 | 4 | TO | 49,369 | 89,502 |
| | 3 | 3 | TO | 60,025 | 85,746 |
| | 4 | 4 | TO | 82,039 | 59,438 |
| | 5 | *2 | TO | 70,246 | 76,229 |
| 30,000 | 1 | 5 | TO | 58,035 | 114,797 |
| | 2 | 3 | TO | 56,441 | 143,617 |
| | 3 | 4 | TO | 60,027 | 150,718 |
| | 4 | 3 | TO | 56,445 | 140,945 |
| | 5 | 4 | TO | 75,841 | 136,768 |
| 223,291 | | 3 | TO | 58,825 | 42,595 |

An asterisk "*" indicates that although a gap is reported by the solver, the solution matches the result obtained by the CBB algorithm.

**Table 10 Results of Hybrid LBBD Algorithm (Algorithm 1) for Different Instance Sizes**

| Size | No. | Algorithm 1 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Gap | Time | #Opt Cuts | #Feas Cuts | #CBB Cuts | #CBB Calls | #CBB Rec | #BC Nodes |
| 1,000 | 1 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 318 |
| | 2 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 378 |
| | 3 | 0 | 0 | 9 | 0 | 0 | 0 | 0 | 430 |
| | 4 | 0 | 0 | 14 | 0 | 0 | 0 | 0 | 469 |
| | 5 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 381 |
| 5,000 | 1 | 0 | 2 | 1,249 | 507 | 183 | 527 | 1,040 | 11,062 |
| | 2 | 0 | 2 | 904 | 542 | 70 | 312 | 495 | 8,738 |
| | 3 | 0 | 1 | 907 | 645 | 43 | 264 | 335 | 9,838 |
| | 4 | 0 | 1 | 766 | 594 | 4 | 195 | 204 | 8,208 |
| | 5 | 0 | 1 | 736 | 540 | 15 | 156 | 175 | 7,954 |
| 10,000 | 1 | 0 | 68 | 8,639 | 4,764 | 6,647 | 2,682 | 11,947 | 68,543 |
| | 2 | 0 | 64 | 7,762 | 4,781 | 6,842 | 2,242 | 7,579 | 67,416 |
| | 3 | 0 | 146 | 10,571 | 9,619 | 3,565 | 6,213 | 24,506 | 116,124 |
| | 4 | 0 | 125 | 9,868 | 8,504 | 3,413 | 5,884 | 23,928 | 108,353 |
| | 5 | 0 | 97 | 8,426 | 7,641 | 2,527 | 4,685 | 16,455 | 98,459 |
| 15,000 | 1 | 0 | 1,782 | 37,201 | 6,523 | 30,033 | 13,640 | 104,373 | 247,888 |
| | 2 | 0 | 2,394 | 39,564 | 7,257 | 34,797 | 14,348 | 93,408 | 283,938 |
| | 3 | 0 | 7 | 577 | 45 | 189 | 61 | 384 | 3,992 |
| | 4 | 0 | 1,819 | 38,479 | 3,742 | 28,496 | 14,400 | 99,106 | 239,752 |
| | 5 | 0 | 412 | 15,025 | 3,024 | 15,139 | 4,967 | 24,285 | 120,737 |
| 20,000 | 1 | *2 | TO | 80,171 | 70,127 | 43,431 | 59,660 | 1,020,734 | 1,755,111 |
| | 2 | *3 | TO | 79,764 | 73,172 | 42,410 | 58,426 | 979,089 | 1,729,008 |
| | 3 | 0 | TO | 80,987 | 70,965 | 44,561 | 60,810 | 913,875 | 1,898,770 |
| | 4 | *4 | TO | 89,610 | 76,794 | 48,648 | 67,473 | 1,123,465 | 1,728,209 |
| | 5 | *2 | TO | 79,347 | 70,623 | 41,677 | 58,268 | 880,261 | 2,011,960 |
| 25,000 | 1 | *3 | TO | 75,160 | 73,049 | 44,231 | 58,623 | 1,677,040 | 1,150,467 |
| | 2 | *3 | TO | 82,335 | 86,002 | 48,200 | 65,717 | 1,360,912 | 1,331,444 |
| | 3 | *2 | TO | 62,595 | 66,641 | 37,470 | 49,364 | 1,185,607 | 1,575,102 |
| | 4 | 3 | TO | 79,725 | 79,898 | 45,936 | 62,952 | 1,294,992 | 1,228,444 |
| | 5 | *1 | TO | 75,146 | 78,327 | 43,072 | 58,636 | 1,310,267 | 1,365,038 |
| 30,000 | 1 | 0 | 752 | 22,737 | 26,965 | 8,336 | 14,763 | 73,165 | 513,167 |
| | 2 | 0 | 532 | 18,620 | 23,116 | 6,683 | 11,490 | 46,786 | 355,877 |
| | 3 | 0 | 1,102 | 20,267 | 26,087 | 7,354 | 12,553 | 52,734 | 646,808 |
| | 4 | 0 | 637 | 19,465 | 23,505 | 6,939 | 12,159 | 53,177 | 379,999 |
| | 5 | 0 | 559 | 18,635 | 24,172 | 6,476 | 11,347 | 47,728 | 303,805 |
| 223,291 | | 2 | TO | 423 | 1 | 201 | 201 | 103,857,204 | 5,678 |

An asterisk "*" indicates that although a gap is reported by the solver, the obtained solution matches the result from the Bron–Kerbosch and CBB algorithms.

**Table 11 Results of Algorithm 1 with Age-Based Side-Constraint for Different Instance Sizes**

| Size | No. | Clique | Best Obj | Gap | Time | #Opt Cuts | #Feas Cuts | #CBB Cuts | #CBB Calls | #CBB Rec | #BC Nodes |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1,000 | 1 | SHMSA | 0.27 | 0 | $<1$ | 6 | 0 | 0 | 0 | 0 | 461 |
| | 2 | NEMD | 0.35 | 0 | $<1$ | 7 | 0 | 0 | 0 | 0 | 402 |
| | 3 | GD | 0.34 | 0 | $<1$ | 7 | 0 | 0 | 0 | 0 | 429 |
| | 4 | NEMD | 0.35 | 0 | 0 | 5 | 0 | 0 | 0 | 0 | 431 |
| | 5 | AFRD, PDX-Unacc | 0.20 | 0 | $<1$ | 3 | 0 | 0 | 0 | 0 | 510 |
| 5,000 | 1 | SHMSA, OLD | 0.47 | 0 | 1 | 76 | 55 | 0 | 2 | 2 | 1,784 |
| | 2 | SHMSA, RCU, E-PO | 0.45 | 0 | 1 | 78 | 56 | 0 | 3 | 3 | 1,926 |
| | 3 | GD, RCU, MD | 0.46 | 0 | 1 | 50 | 27 | 0 | 0 | 0 | 1,422 |
| | 4 | GD, DOA, MD | 0.52 | 0 | 0 | 35 | 22 | 0 | 0 | 0 | 1,340 |
| | 5 | SHMSA, OLD | 0.46 | 0 | 1 | 38 | 14 | 0 | 0 | 0 | 1,419 |
| 10,000 | 1 | OSULD, HepF, PDX-Unacc, FED | 0.58 | 0 | 9 | 273 | 197 | 13 | 48 | 73 | 4,484 |
| | 2 | OSULD, HepF | 0.61 | 0 | 7 | 224 | 142 | 8 | 32 | 49 | 3,651 |
| | 3 | OLD, ONSD, NEMD | 0.56 | 0 | 5 | 246 | 277 | 3 | 31 | 35 | 4,454 |
| | 4 | E-AMD, MD, NEMD | 0.54 | 0 | 4 | 249 | 238 | 8 | 45 | 54 | 4,428 |
| | 5 | AD, DOA, RCU, MD | 0.47 | 0 | 3 | 152 | 94 | 5 | 15 | 24 | 2,682 |
| 15,000 | 1 | PDX-Unacc, HepF | 0.57 | 0 | 41 | 646 | 576 | 88 | 192 | 452 | 10,989 |
| | 2 | E-AMD, MD, NORCP | 0.63 | 0 | 41 | 742 | 769 | 100 | 200 | 479 | 13,416 |
| | 3 | AFRD, PDX-Unacc, CNTN | 0.62 | 0 | 41 | 742 | 692 | 114 | 211 | 502 | 12,440 |
| | 4 | Hepatitis, OLD, FD | 0.67 | 0 | 45 | 651 | 622 | 78 | 159 | 332 | 11,402 |
| | 5 | DCP, AlcD, FD | 0.59 | 0 | 12 | 359 | 320 | 27 | 66 | 131 | 6,142 |
| 20,000 | 1 | Hepatitis, OLD, ARenF | 0.64 | 0 | 99 | 1,307 | 1,306 | 249 | 464 | 1,597 | 22,147 |
| | 2 | DSD, OLD | 0.63 | 0 | 131 | 1,788 | 1,921 | 346 | 687 | 2,415 | 31,567 |
| | 3 | PDX-Unacc, NV, CNTN | 0.67 | 0 | 89 | 1,415 | 1,333 | 281 | 544 | 1,674 | 22,190 |
| | 4 | Hepatitis, OLD, FD | 0.65 | 0 | 122 | 1,715 | 1,854 | 350 | 680 | 2,025 | 30,441 |
| | 5 | AFRD, PDX-Unacc, CNTN | 0.76 | 0 | 126 | 1,106 | 948 | 224 | 442 | 1,276 | 19,946 |
| 25,000 | 1 | RCU, FevUO, OLD | 0.68 | 0 | 302 | 2,252 | 2,602 | 511 | 936 | 3,524 | 43,909 |
| | 2 | ONSD, IntI, MD | 0.67 | 0 | 397 | 3,537 | 3,848 | 992 | 1,712 | 8,212 | 64,121 |
| | 3 | PIA, OLD | 0.69 | 0 | 262 | 2,530 | 2,551 | 665 | 1,132 | 4,946 | 42,931 |
| | 4 | Hepatitis, OLD, ARenF | 0.68 | 0 | 267 | 2,984 | 3,081 | 803 | 1,418 | 5,179 | 51,320 |
| | 5 | DCP, AlcD, OLD, ARenF | 0.70 | 0 | 226 | 2,321 | 2,101 | 559 | 1,062 | 3,769 | 39,752 |
| 30,000 | 1 | RCU, DOA, AlcD | 0.60 | 0 | 29 | 766 | 1,322 | 31 | 178 | 236 | 23,009 |
| | 2 | AlcD, OLD | 0.65 | 0 | 18 | 536 | 999 | 10 | 103 | 117 | 13,311 |
| | 3 | PS | 0.61 | 0 | 24 | 505 | 924 | 23 | 110 | 138 | 20,493 |
| | 4 | AlcD, OLD | 0.66 | 0 | 21 | 579 | 945 | 15 | 135 | 154 | 14,045 |
| | 5 | AlcD, OLD | 0.64 | 0 | 23 | 535 | 889 | 21 | 116 | 142 | 14,355 |
| 223,291 | | - | - | - | TO | 11,263 | 35 | 5,571 | 5,593 | 37,466,111 | 51,196 |

Disease names are abbreviated for space reasons; full names are provided in Appendix D.

A dash "-" indicates that no value was reported by the solver for that metric.

**Table 12** **Values of $a_{60}$ for Diseases Appearing in 25,000-Patient-Sample Solutions**

| No. | 1 | 2 | 3 | 4 | 5 | Avg | Range |
|---|---|---|---|---|---|---|---|
| AlcD | 0.23 | 0.21 | 0.22 | 0.22 | 0.22 | 0.22 | [0.21, 0.23] |
| ARenF | 0.72 | 0.70 | 0.70 | 0.71 | 0.71 | 0.71 | [0.70, 0.72] |
| AURF | 0.68 | 0.68 | 0.68 | 0.68 | 0.67 | 0.68 | [0.67, 0.68] |
| CA | 0.78 | 0.79 | 0.79 | 0.77 | 0.78 | 0.78 | [0.77, 0.79] |
| CoagHD | 0.61 | 0.59 | 0.59 | 0.59 | 0.60 | 0.60 | [0.59, 0.61] |
| DCP | 0.56 | 0.56 | 0.57 | 0.57 | 0.58 | 0.57 | [0.56, 0.58] |
| DLM | 0.73 | 0.73 | 0.73 | 0.72 | 0.73 | 0.73 | [0.72, 0.73] |
| FED | 0.62 | 0.62 | 0.62 | 0.61 | 0.61 | 0.62 | [0.61, 0.62] |
| FevUO | 0.44 | 0.47 | 0.47 | 0.47 | 0.45 | 0.46 | [0.44, 0.47] |
| Hepatitis | 0.29 | 0.27 | 0.25 | 0.28 | 0.29 | 0.28 | [0.25, 0.29] |
| HF | 0.79 | 0.79 | 0.79 | 0.79 | 0.80 | 0.79 | [0.79, 0.80] |
| IntI | 0.58 | 0.54 | 0.56 | 0.56 | 0.57 | 0.56 | [0.54, 0.58] |
| MD | 0.39 | 0.38 | 0.39 | 0.38 | 0.38 | 0.38 | [0.38, 0.39] |
| NCD | 0.90 | 0.88 | 0.88 | 0.85 | 0.87 | 0.88 | [0.85, 0.90] |
| OACE | 0.72 | 0.72 | 0.72 | 0.71 | 0.72 | 0.72 | [0.71, 0.72] |
| OLD | 0.48 | 0.47 | 0.47 | 0.47 | 0.48 | 0.47 | [0.47, 0.48] |
| ONSD | 0.57 | 0.56 | 0.56 | 0.57 | 0.57 | 0.56 | [0.56, 0.57] |
| OSNSD | 0.68 | 0.67 | 0.67 | 0.69 | 0.70 | 0.68 | [0.67, 0.70] |
| PDX-Unacc | 0.51 | 0.51 | 0.51 | 0.50 | 0.50 | 0.51 | [0.50, 0.51] |
| PIA | 0.52 | 0.47 | 0.48 | 0.41 | 0.48 | 0.47 | [0.41, 0.52] |
| PNA | 0.63 | 0.62 | 0.62 | 0.62 | 0.63 | 0.62 | [0.62, 0.63] |
| RCU | 0.56 | 0.54 | 0.55 | 0.56 | 0.55 | 0.55 | [0.54, 0.56] |
| RF | 0.66 | 0.66 | 0.66 | 0.65 | 0.66 | 0.66 | [0.65, 0.66] |
| SEPT | 0.62 | 0.61 | 0.61 | 0.61 | 0.61 | 0.61 | [0.61, 0.62] |
| SHK | 0.68 | 0.67 | 0.67 | 0.68 | 0.67 | 0.67 | [0.67, 0.68] |

Disease names are abbreviated for space reasons; full names are provided in Appendix D.

**Table 13** **Values of $a_{60}$ for Diseases Appearing in 30,000-Patient-Sample Solutions**

| No. | 1 | 2 | 3 | 4 | 5 | Avg | Range |
|---|---|---|---|---|---|---|---|
| AlcD | 0.22 | 0.20 | 0.22 | 0.21 | 0.21 | 0.21 | [0.20, 0.22] |
| ARF | 0.62 | 0.62 | 0.64 | 0.62 | 0.66 | 0.63 | [0.62, 0.66] |
| CA | 0.80 | 0.81 | 0.81 | 0.81 | 0.80 | 0.81 | [0.80, 0.81] |
| CardArr | 0.80 | 0.78 | 0.79 | 0.79 | 0.79 | 0.79 | [0.78, 0.80] |
| CHF-NH | 0.83 | 0.82 | 0.82 | 0.82 | 0.83 | 0.83 | [0.82, 0.83] |
| CUS | 0.73 | 0.74 | 0.70 | 0.73 | 0.71 | 0.72 | [0.70, 0.74] |
| DOA | 0.59 | 0.57 | 0.59 | 0.57 | 0.58 | 0.58 | [0.57, 0.59] |
| GD | 0.57 | 0.59 | 0.59 | 0.58 | 0.58 | 0.58 | [0.57, 0.59] |
| HCSH | 0.79 | 0.80 | 0.80 | 0.79 | 0.80 | 0.80 | [0.79, 0.80] |
| OA | 0.62 | 0.63 | 0.63 | 0.62 | 0.62 | 0.62 | [0.62, 0.63] |
| OLD | 0.50 | 0.50 | 0.52 | 0.51 | 0.50 | 0.51 | [0.50, 0.52] |
| ONSD | 0.56 | 0.58 | 0.56 | 0.57 | 0.58 | 0.57 | [0.56, 0.58] |
| PS | 0.49 | 0.48 | 0.49 | 0.48 | 0.48 | 0.48 | [0.48, 0.49] |
| RCU | 0.55 | 0.55 | 0.57 | 0.56 | 0.56 | 0.56 | [0.55, 0.57] |
| SHK | 0.66 | 0.67 | 0.69 | 0.67 | 0.68 | 0.67 | [0.66, 0.69] |

Disease names are abbreviated for space reasons; full names are provided in Appendix D.

## Appendix D: Disease Abbreviations

Table 14 lists the full disease names along with the abbreviations used in the solutions.

**Table 14 Disease Abbreviations and Full Names**

| Abbreviation | Description |
|---|---|
| AD | Anxiety disorders |
| AFD | Abnormal findings without diagnosis |
| AFRD | Anxiety and fear-related disorders |
| AlcD | Alcohol-related disorders |
| ARF | Adult respiratory failure |
| ARenF | Acute renal failure |
| AURF | Acute and unspecified renal failure |
| CA | Coronary atherosclerosis and other heart disease |
| CardArr | Cardiac arrhythmia |
| CD | Cardiac dysrhythmias |
| CHF-NH | Congestive heart failure; nonhypertensive |
| CKD | Chronic kidney disease |
| CNTN | Conditions due to neoplasm or the treatment of neoplasm |
| CoagHD | Coagulation and hemorrhagic disorders |
| COPD-BE | Chronic obstructive pulmonary disease and bronchiectasis |
| CUS | Chronic ulcer of skin |
| DCP | Other diagnostic cardiovascular procedures |
| DLM | Disorders of lipid metabolism |
| DOA | Deficiency and other anemia |
| DSD | Other disorders of stomach and duodenum |
| E-AMD | E Codes: Adverse effects of medical drugs |
| E-PO | E Codes: Place of occurrence |
| FD | Fluid disorders |
| FED | Fluid and electrolyte disorders |
| FevUO | Fever of unknown origin |
| GD | Other gastrointestinal disorders |
| HCSH | Hypertension with complications and secondary hypertension |
| HF | Heart failure |
| HepF | Hepatic failure |
| Hepatitis | Hepatitis |
| IntI | Intestinal infection |
| MALN | Malnutrition |
| MD | Mood disorders |
| NCD | Neurocognitive disorders |
| NEMD | Other nutritional; endocrine; and metabolic disorders |
| NORCP | Other non-OR therapeutic cardiovascular procedures |
| NV | Nausea and vomiting |
| OA | Other aftercare |
| OACE | Other aftercare encounter |
| OGS | Other general signs and symptoms |
| OLD | Other liver diseases |
| ONSD | Other nervous system disorders |
| OSNSD | Other specified nervous system disorders |
| OSULD | Other specified and unspecified liver disease |
| PDX | Code is unacceptable as a principal diagnosis (only used for the inpatient default CCSR) |
| PDX-Unacc | Other PDX-Unacceptable |
| PIA | Peritonitis and intestinal abscess |
| PNA | Pneumonia (except that caused by tuberculosis) |
| PS | Procedures on spleen |
| RCU | Residual codes; unclassified |
| RF | Respiratory failure; insufficiency; arrest |
| RSS | Respiratory signs and symptoms |
| SEPT | Septicemia |
| SEPTL | Septicemia (except in labor) |
| SHK | Shock |
| SHMSA | Screening and history of mental health and substance abuse codes |
| SM | Secondary malignancies |