Optimal Trauma Care Network Redesign with Government Subsidy: A Bilevel Integer Programming Approach

Trauma presents a prominent health problem worldwide. However, trauma centers are often clustered in urban areas and sparsely located in rural areas. The geographic maldistribution of trauma centers leads to system-related mistriage errors. While some local governments offer subsidy to incentivize the affiliated hospital group to redesign the trauma care network, the approach is ad hoc. To address this issue, we propose a bilevel integer programming model to investigate the subsidized trauma care network redesign problem, which considers the government as the leader and the hospital group as the follower. To solve the resultant problem efficiently, we propose a branching idea to exclude additional infeasible solutions and suboptimal solutions, in turn speeding up the branch-and-bound algorithm. In a case study, we redesign a trauma care network in the midwestern area of the U.S. based on closed-form approximate functions of system-related mistriage errors. The results show that the optimal network redesign redistributes the network by slightly reducing the number of trauma centers to relieve the crowded trauma care resource, and achieves an overall improvement of about 11% over the original network.

Key words: emergency care, network design, facility location optimization, bilevel integer programming, government subsidy

1. Background and Introduction

Trauma refers to severe physical injuries of sudden onset caused by violence or accident, which may lead to death if definitive care is not administered in a timely fashion. Trauma is a serious public health problem with significant social and economic burden. In the United States (U.S.), it is the #1 cause of death among citizens younger than 45 years old and ranks third overall across all ages (Rhee et al. 2014), accounting for nearly 200,000 deaths annually (NTI 2016).

American College of Surgeons (ACS) designates trauma care facilities nationwide from Level 1 (L1) to level 5 (L5), based on their level of trauma care specialty (ATS 2016), and availability of specialized trauma care resource (emergency department beds, capital equipment, and specialty care staff). Both L1/L2 designated trauma care facilities provide nearly the same care to severely-injured patients (often life-threatening). They are required to have 24/7 in-house coverage and prompt availability of care in surgical specialities such as orthopedic, neuro, plastic, and oral and maxillofacial. In contrast, L3-L5 designated trauma care facilities provide only a subset of the services mentioned above, house limited 24/7 trauma staff, and often serve as the backup trauma

care facilities for severely injured patients in areas without an L1/L2 designated facility (ATS 2016). In this paper, we refer to L1/L2 designated trauma care facilities as trauma centers (in short, TCs) and L3-L5 designated trauma care facilities, along with other community hospitals, as non-trauma centers (in short, NTCs). Normally, to become a TC, any facility must have a sufficient volume of specialized trauma care resource.

When emergency medical service (EMS) staff arrive at the scene of a trauma incident, they perform preliminary assessment on the victim and make a field-triage decision to determine where to transport the victim. Typically, the staff consider two crucial factors, namely injury severity (based on several clinical factors, e.g., respiratory rate, blood pressure, and Glasgow Coma Score), and transport time/distance (i.e., whether a TC is within the geographic proximity of the incident). Besides the injury severity, it is well evident that patient's prehospital survival and care outcomes are strongly correlated to the time it takes for the patient to access appropriate trauma care (Brown et al. 2016). Ideally, severely injured patients are transported to TCs for specialized care, whereas non-severely injured patients are transported to NTCs so as not to compete for precious specialized care resources with severely injured ones.

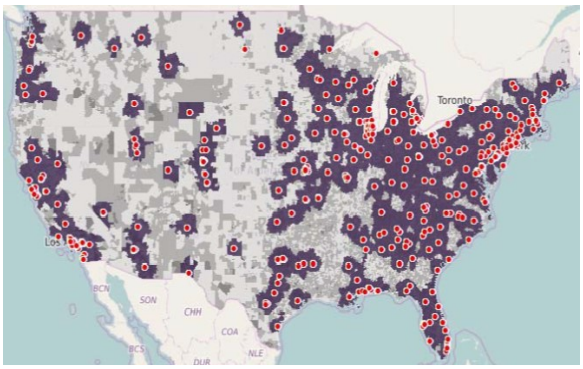


Figure 1 Geographic maldistribution in trauma care in 2010 (Red dots = TCs, dark shade = 60-min coverage by some TC, light shade = distribution of U.S. population).

However, it is often not the case, for which a main reason is the geographic maldistribution of trauma care facilities (Brown et al. 2016). At present, TCs are typically clustered in urban areas, while distributed sparsely (sometimes absent) in socioeconomically disadvantaged (often rural and under-funded) areas. Figure 1 shows the distribution of nearly 250 TCs in the U.S. in 2005, which covered only 30% of land within the golden hour (60 minutes) via ambulance and helicopter (Branas et al. 2005, Carr and Branas 2010). This coverage leaves no access to a trauma center for nearly 45 million Americans within the golden hour, according to the Centers for Disease Control and Prevention (ATS 2016). The geographic maldistribution has caused serious implications both socioeconomically and in terms of health outcomes in U.S. (Brown et al. 2016). Since

2012, there has been further an upswing (over 117 TCs opened and existing hospitals upgraded) largely into metropolitan areas (Galewitz 2012). This expansion of TCs, has intensified the geographic maldistribution and created huge outcry in many under-privileged areas. For example, six TCs were built along the Atlantic coast of Florida, instead of three that were really needed as per Florida's state health department (Chang 2016), which presents a huge contrast to the limited specialized trauma care resource in Florida Panhandle, a more rural region of Florida.

Due to lack of TCs within a critical time threshold of the incident location, EMS staff are forced to transport a severely injured patient to nearby NTCs, which is referred to as a system-related under-triage (srUT) error. Similarly, with an excess of TCs in the vicinity of the incident location, the EMS may be induced to transport a non-severely injured patient to one of those TCs (instead of an NTC), which is referred to as a system-related over-triage (srOT) error (Newgard et al. 2013). On one hand, srUT errors often occur in areas that TCs are far away and patients cannot receive timely and appropriate specialized care. Thus, these errors lead to increased risk of short- or long-term disability or even mortality as severely injured patients cannot get timely and appropriate emergency care. Severely injured patients taken to TCs would have a 25% higher survival rate than taken to NTCs (MacKenzie et al. 2006). On the other hand, srOT errors often occur in areas that TCs are clustered and patients may be transported to nearby TCs even if the injuries are not severe. Thus, these errors lead to unnecessary spending on specialized trauma care to non-severely injured patients at TCs. On average, trauma care payment is 41% higher at TCs than at NTCs for identical treatments (Newgard et al. 2013). The higher-level payment is mainly due to a fixed trauma activation fee that each TC charges to its patients to recoup its high cost of 24/7 on-call trauma care staff and specialized medical equipment. This fee can range \$5,000 – \$27,000, depending on the state and region (Singer 2012). See Appendix I for two possible real-world scenarios of system-related mistriage errors.

Given this, from the viewpoint of social welfare (government's objective), balancing the distribution of trauma care resource in the area is pivotal. This may include upgrading some NTCs to TCs in areas with scarce trauma care resource, while downgrading some TCs to NTCs in areas with over-capacity trauma care resource. In contrast, from the viewpoint of financial viability (hospital group's objective), the change in the volume of patients assigned to TCs may impact the revenues generated to offset high fixed and operating costs of TCs (Singer 2012, ATS 2016). This is why some governments provide subsidy to the hospital group to strike a compromise between social welfare and hospital group's financial viability.

In this paper, we study a subsidized network redesign problem (SNrDP), aiming to reduce the system-related mistriage errors of an existing trauma network. This subsidized network redesign involves two main stakeholders, namely the government and the hospital group. While many

state/regional governments have limited authority to enforce ideal care delivery from a societal perspective, they can clearly influence the network design by providing subsidy to the hospital group. The government's objective is to reduce the overall harm caused by srOT and srUT errors without providing a large amount of subsidy. In contrast, the hospital group intends to maximize the profit by optimally locating TCs and specifying the coverage zone of each TC. To analyze the interplay between them, we formulate the SNrDP as a bilevel integer programming model in which the government is the leader and the hospital group is the follower. We consider the optimistic formulation of the bilevel programming. That is, when the lower-level problem possesses multiple optimal solutions, the leader gets to pick among them so as to achieve the best upper-level objective.

The proposed bilevel integer programming model is difficult to be solved exactly for large-scale, real-world instances. In general, bilevel integer programs are known to be NP-hard even when the leader's and follower's problems are both linear programs (Jeroslow 1985). Further, different from almost all bilevel integer programming problems studied previously in the operations research literature, our problem involves a large-scale combinatorial optimization model at the lower level. To address the above additional computational challenge, we propose a branching idea to superimpose on standard branching schemes so as to exclude more suboptimal solutions and infeasible solutions at each branch-and-bound iteration. We verify our model and test our solution method with a well-designed case study of redesigning a trauma care network in the midwest of the United States. Our case study suggests around 11% overall improvement (that is a decreased quantification value of the mistriage errors) can be achieved from the original network by redistributing the existing TCs. This would require modest dispersion throughout the area. We also conduct sensitive analysis on several key model parameters to assess their correlations with the network redesign.

Our main contributions are three-fold. First, the proposed leader-follower bilevel programming model can be widely applied to discrete location optimization with subsidy agreement between the leader and the follower. Arising in many public sector operations research application contexts, it is important to establish an effective public-private partnership (PPP) that addresses the geographic demand-supply imbalance in a service network with multi-class customers (e.g., severely injured and non-severely injured patients). Second, we design an exact branch-and-bound algorithm, which tailors a previously self-developed branching rule (Liu et al. 2021) to our specific formulation with the upper-level decision being one-dimensional. Note that the previously developed branching rule cannot ensure the exactness of the algorithm if there are multiple optimal solutions in the lower-level problem. Our tailored branching idea can be embedded into any standard branch-and-bound framework for the BIP problem so as to speed up the algorithm. Lastly, our real-world case study

is expected to be insightful to make good decisions on subsidy budgeting and care network design, which can help reduce mistriage errors of trauma care in the target area.

The remainder of this paper is organized as follows. In Section 2, we provide a comprehensive review of the relevant literature. In Section 3, we present a bilevel integer programming model for the SNrDP. In Section 4, we propose an exact branch-and-bound algorithm. In Section 5, we report a case study based on real data and make policy recommendations. We draw concluding remarks and outline future research in Section 6.

2. Literature Review

2.1. Trauma care outcomes research

Relevant studies from trauma care outcomes research include those on prehospital trauma triage and those on trauma care network design. In the former category, several studies focus on prehospital triage outcomes, including Newgard et al. (2013), Carr et al. (2016), Parikh et al. (2017), and Parikh et al. (2019). These studies apply statistical methods to investigate the impact of triage errors on medical cost, mortality rate and so on. A variety of methods and scores exist in the current literature for prehospital trauma triage, including Baxt et al. (1990), Mackersie (2006), Newgard et al. (2011), Sasser et al. (2011), and Jones et al. (2016). These studies provide us a good understanding of triage protocols in EMS decision-making and subject-matter support to out development of the prehospital triage outcome simulation.

To design trauma care networks, the aspect of geographical maldistribution of TCs has garnered tremendous attention in the medical literature. These studies give rise to the use of standard operations research techniques to selected aspects of the trauma care network design problem. For instance, Branas et al. (2005) considered a single centralized network designer and used a heuristic method to locate trauma facilities. In another single-planner study, Branas et al. (2013) evaluated the marginal impact of adding one or two TCs and helicopter depots within 45/60-minute access to these centers by enumerating all feasible location solutions. Jansen et al. (2014) and Jansen et al. (2015) conducted geospatial evaluation studies on designing trauma care networks in Scotland, and identified Pareto-optimal solutions to the two conflicting objectives, i.e., travel time and under-triage errors. Carr et al. (2016) used simulation to estimate the distribution of trauma incidents, mortality rates, and care resource utilization for disaster preparedness in 25 largest U.S. cities. On the operations research side, there is research on ambulance location and operations decision problems for improving the performance of a trauma care system, including Knight et al. (2012), Sudtachat et al. (2016), Enayati et al. (2018) and Bélanger et al. (2020). None of these studies captured the effect of government subsidy on network designer's decisions, nor the tradeoff between triage errors and the subsidy, all of which are critical to appropriately addressing the trauma care network design problem in the U.S.

2.2. Healthcare facility location analysis

Our paper is also related to the healthcare facility location analysis research in the operations research literature. Because of the large volume of the literature in this area, we review only the most relevant papers and refer readers to Daskin and Dean (2004), Li et al. (2011), and Ahmadi-Javid et al. (2017) for comprehensive reviews. Healthcare facility location problems with a single planner are well-studied; see, Rahman and Smith (2000), Verter and Lapierre (2002), Harper et al. (2005), Jia et al. (2007), Griffin et al. (2008), Vidyarthi and Jayaswal (2014), and Mestre et al. (2015). There are several papers studying the location of TCs; see, Branas et al. (2000), Branas and Revelle (2001), Côté et al. (2007), Syam and Côté (2010), Erdemir et al. (2010), Cho et al. (2014), and Lee and Jang (2018). Branas et al. (2000), Cho et al. (2014), and Lee and Jang (2018) studied the optimization problem of jointly locating TCs and associated helicopter (i.e., air ambulance) platforms and depots. Erdemir et al. (2010) considered the possibility of deploying both ambulances and helicopters to transfer patients to TCs when the scene of an incident does not have a suitable nearby area for a helicopter to safely land. Since sustaining TC operations is expensive, these studies either dealt with cost minimization of locating and operating TCs from the viewpoint of a hospital network, or care availability/accessibility maximization under budget constraint from the viewpoint of the government. None of theses studies considered modeling of the above conflicting interests simultaneously, nor considered a trauma care network composed of both TCs and NTCs.

The literature on healthcare facility location remains scarce when it comes to concerning multiple decision-makers with conflicting interests. A few papers present bilevel optimization models for the system planner in the public health/humanitarian sector, which only take into account service recipients' behaviors through some form of equilibrium constraint. Thus these models can be equivalently reduced to a single planner's models (Gutjahr and Dzubur 2016, Zhang et al. 2010). To the best of our knowledge, no healthcare facility location research addresses the cases where the system planner and the service provider play a leader-follower game.

Outside the literature on healthcare facility location, a series of studies investigate a defender-attacker facility location problem, where two decision-makers play a leader-follower game (Scaparra and Church 2008, Küçükaydin et al. 2011, Liberatore et al. 2012, Keçici et al. 2012, Aksen and Aras 2012, Aksen et al. 2014, Ghaffarinasab and Motallebzadeh 2018, Ghaffarinasab and Atayi 2018, Haywood et al. 2022). In these studies, the leader (defender) aims to maximize the coverage of customer zones by relocating the facilities, while the follower (attacker) seeks to maximize the destroy of the customer service system. These papers generally propose a variety of heuristic approaches to solve the resultant discrete bilevel programming problems, and to achieve well-performed solutions which though cannot ensure the bilevel feasiblity. In addition, Aksen et al. (2009) and Bhadury

and Eiselt (2012) studied a subsidy optimization problem with multiple decision-makers. Aksen et al. (2009) formulated two bilevel programming models describing subsidy agreement between the government and the company engaged in collection and recovery operations. Bhadury and Eiselt (2012) introduced a three-level model for a subsidized location optimization problem where the regional planner offers a subsidy to the firm and the firm establishes a number of distribution centers at different locations. These two papers solved their models with heuristic methods. Moreover, several researchers investigated the PPP between a government and a private party with the objective of delivering public assets and/or services (Lavlinskii et al. 2021, 2018, 2015, Rodríguez 2020). The interactions in a PPP can be viewed as a leader-follower game and formulated as a bilevel programming problem. These papers generally proposed effective solution algorithms based on metaheuristics and local search, or incorporated recent branch-and-bound algorithms found in the literature to solve simulated instances.

2.3. Bilevel integer programming

In recent years, there is a growing literature on discrete bilevel programming. Here we present a review of exact algorithms within the branch-and-bound framework for bilevel mixed integer linear programming (BMILP) and bilevel integer linear programming (BILP), which are mostly related to our research. An early study is found in Moore and Bard (1990), which proposed the first exact algorithm for BMILP. Their algorithm is shown to converge in two cases: either when all leader variables are integer, or when the follower subproblem is a linear program. Following their idea, DeNegre and Ralphs (2009) used the cutting plane technique to propose a branch-and-cut algorithm, and consequently encountered fewer nodes in the branch-and-bound tree than Moore and Bard (1990). Xu and Wang (2014) designed an exact branch-and-bound algorithm for BMILP with bounded and integral assumptions on the upper-level variables. They tested their algorithm on instances with up to 920 variables and 368 constraints. Wang and Xu (2017) integrated the branch-and-bound framework with the cutting-plane technique to present a so-called watermelon algorithm for BILP, which relies on no additional simplifying assumptions. Based on the work of Xu and Wang (2014), Liu et al. (2021) proposed an enhanced branch-and-bound algorithm which takes advantage of the property of the lower-level problem. Their algorithm behaves well especially for BILP problem with complex large-sized lower-level problem. Fischetti et al. (2016, 2018) employed a group of intersection cuts valid to BMILP under some mild assumptions to propose a branch-and-cut algorithm and developed a new family of cuts for BMILP. Further, Fischetti et al. (2017) extended the algorithm in Fischetti et al. (2018) by suggesting new intersection cuts (e.g., hypercube intersection cut), which allows for nonlinear terms appearing in both constraints and objective functions. They tested their algorithm on more than 800 instances from the literature and

demonstrated the superiority of their algorithm. Other algorithms using cutting-plane techniques can be found in Caramia and Mari (2015), Hemmati and Smith (2016), Zhang and Özaltın (2017), and Tahernejad et al. (2020). Besides using branch and bound, a few other solution methods involve novel reformulations and decomposition strategies (Zeng and An 2014, Yue et al. 2019), which include Benders decomposition (Saharidis and Ierapetritou 2009) and parametric programming (Fáisca et al. 2007, Avraamidou and Pistikopoulos 2019). However, these papers only reported computational experiments on small-sized instances.

There are multiple approaches to solve real-world BMILP problems, e.g., Caramia and Mari (2016) and Zare et al. (2019) for facility location problems; Dempe et al. (2005), Kalashnikov and Ríos-Mercado (2006), and Dempe et al. (2011) for natural gas regulation. Caramia and Mari (2016) proposed an algorithm based on decomposition that is similar to the algorithm proposed by Saharidis and Ierapetritou (2009), but the algorithm was further adapted for the control of the leader to cope with the integrality imposed on the variables and the bilevel structure in the facility location problem. Recently, Zare et al. (2019) proposed two strong-duality-based reformulations of the BMILP problems with continuous linear lower-level variables. They tested their approaches on various classes of BMILP instances, including the bilevel facility location instance containing 40 facilities and 240 products at most. Dempe et al. (2005), Kalashnikov and Ríos-Mercado (2006), and Dempe et al. (2011) formulated their bilevel programming problems to equivalent BMILP models. They designed their algorithms based on a penalty-function approach and tested their algorithms on real-world instances with dimensions up to 1000.

Another special case is a family of min-max bilevel problems, where the leader seeks to minimize the follower's objective. Brotcorne et al. (2013) proposed an exact algorithm with dynamic programming and branch-and-bound technique to solve the bilevel knapsack problem (BKP). Their algorithm can solve the BKP with multi-dimensional variables in both the upper and the lower levels, and the lower-level model containing only one constraint. Tang et al. (2016) proposed three generic solution algorithms and required leader variables to be binary, whereas the follower can be a general mixed-integer linear program. Fischetti et al. (2019) presented an exact branch-and-cut algorithm for two-person interdiction games under the assumption that feasible solutions of the follower problem satisfy a certain monotonicity property. Tanınmış et al. (2022) improved the x-space algorithm proposed by Tang et al. (2016) for a recent min–max bilevel optimization problem that arises in the context of reducing the misinformation spread in social networks. Their algorithm even compared favorably with the algorithm in Fischetti et al. (2017) developed for mixed-integer bilevel linear programs. The above studies were only tested on randomly generated BKP instances with dimensions up to 500.

In this paper, we propose a branching idea for BIP problems with the upper-level decision being one-dimensional, so as to augment the standard branching rule in a branch-and-bound framework. This branching idea can further carve out additional infeasible solutions and suboptimal solutions from the search space in each iteration of the branch-and-bound algorithm.

Table 1 Notation

Indices
I: set of demand nodes (e.g., zipcode, town) where traumatic injury incidents occur
J: set of hospital locations in the trauma care network
Model Parameters & Coefficients
Upper-level:
C^T: unit specialty care cost for trauma patients at TC
C^{NT}: unit care cost for trauma patients at NTC
$D_{ij} = 1$, if demand node i is covered by the TC at location j within a coverage-requirement-related distance of d_0; 0, otherwise
γ: health hazard related cost due to delayed specialty care for a severely injured patient (result of an srUT error)
σ: weighting coefficient on government's subsidy over its healthcare spending
δ: coverage required by the government (i.e., proportion of demand nodes; $0 < \delta \leq 1$)
z_i: intermediate variable which is binary
Lower-level:
$H_j = 1$, if there is a TC originally at location j; 0, otherwise
T_i: annual number of trauma incidents at demand node i
R_{min}: annual patient volume required to maintain the operations of a TC
C_j^U: annual operating cost surplus resulting from upgrading the NTC at location j to a TC
C_j^D: annual operating cost saving resulting from downgrading the TC at location j to an NTC
R_{ij}^T: reimbursement payment for providing specialty care to a patient from node i to the TC at location j
$\bar{D}_{ij} = 1$, if demand node i is covered by the TC at location j within a demand-node-related distance of $\bar{d}_i$; 0, otherwise
S_j: annualized monetary compensation for upgrading the NTC at location j (requested *a priori* by the hospital group)
Decision variables
s_c: annualized total amount of subsidy committed by the government to the NTC upgrading
$x_{ij} = 1$, if patients from node i are assigned to TC location j for speciality care; 0, otherwise
$y_j = 1$, if a TC is at location j; 0, if an NTC is at the location

3. Problem Formulation

In this section, we formulate the SNrDP with a bilevel integer programming (BIP) model. Our model involves two distinct decision-makers at two levels, namely the government at the upper level and the hospital group at the lower level. As a leader, the government's goal is to improve the performance of the trauma care network without incurring a significant amount of subsidy for upgrading NTCs. We quantify the network performance with the negative effects caused by over-triage (srOT) and under-triage (srUT) errors. As mentioned in Section 1, srOT and srUT

errors are affected by the network design. Here we introduce two real-valued functions $O(y)$ and $U(y)$ to quantify srOT and srUT errors with respect to any given network design y. In Section 5, we present in detail how we perform data-driven modeling for $O(y)$ and $U(y)$ with respect to a field-triage protocol suggested in the medical literature. As a follower, the goal of the hospital group, in turn, is to maximize the marginal profit through the network redesign. We next introduce the BIP model notation (Table 1) and present the model.

Upper-level model (Government's problem):

$$\min_{s_c,z} \quad Gov(s_c,y) \triangleq (C^T - C^{NT})O(y) + \gamma U(y) + \sigma s_c, \tag{1}$$

$$\text{s.t.} \quad \sum_{j\in J} D_{ij} y_j \geq z_i, \quad i \in I, \tag{2}$$

$$\sum_{i\in I} z_i \geq \delta |I|, \tag{3}$$

$$z_i \in \{0,1\}, \quad i \in I, \tag{4}$$

$$s_c \in \mathbb{Z}_+, \tag{5}$$

where $z = \{z_i\}_{i\in I}$ are intermediate variables, and $y = \{y_j\}_{j\in J}$ is optimal to the lower-level problem denoted as $\mathcal{L}(s_c)$ for a fixed s_c.

Lower-level model (Hospital group's problem) $\mathcal{L}(s_c)$:

$$\begin{aligned}\max_{\tilde{x},\tilde{y}} \quad Hos(\tilde{x},\tilde{y}) \triangleq \; & -\sum_{j\in J}(C_j^U - S_j)(1-H_j)\tilde{y}_j + \sum_{j\in J} C_j^D H_j (1-\tilde{y}_j) \\ & + \sum_{i\in I}\sum_{j\in J} T_i \tilde{x}_{ij}(R_{ij}^T - C^T) \end{aligned} \tag{6}$$

$$\text{s.t.} \quad \tilde{x}_{ij} \leq \bar{D}_{ij}\tilde{y}_j, \quad i \in I, j \in J, \tag{7}$$

$$\sum_{j\in J} \tilde{x}_{ij} = 1, \quad i \in I, \tag{8}$$

$$\sum_{j\in J} S_j(1-H_j)\tilde{y}_j \leq s_c, \tag{9}$$

$$\sum_{i\in I} T_i \tilde{x}_{ij} \geq R_{min}\tilde{y}_j, \quad j \in J, \tag{10}$$

$$\tilde{x}_{ij}, \tilde{y}_j \in \{0,1\}, \quad i \in I, j \in J, \tag{11}$$

where $\tilde{x} = \{\tilde{x}_{ij}\}_{i\in I, j\in J}$, $\tilde{y} = \{\tilde{y}_j\}_{j\in J}$.

In the upper-level model, the government determines an annualized amount of subsidy denoted by s_c. The government's objective function, $Gov(s_c, y)$, contains three parts: (a) additional treatment cost due to srOT errors; (b) health hazard cost related to srUT errors; and (c) subsidy. For part (a), when an srOT error occurs, a non-severely injured patient is taken to TC and operated on with specialty care. Thus the additional unit treatment cost is the differential between the unit treatment

costs at a TC and an NTC, i.e., $C^T - C^{NT}$. For part (b), when an srUT error occurs, a severely injured patient is mistakenly taken to NTC, at which the diagnosis shows an srUT error. Thus the eminent health hazard prompts an immediate TC transport but it will still cause elevated morbidity due to delayed treatment. These sequelae incur additional costs, i.e., γ is the additional costs per UT error. Note that these two cost terms (a) and (b) are regarded as system-wide spending items to the government, especially if only publicly funded beneficiaries (i.e., Medicaid and Medicare patients) are considered. Thus the two cost terms are combined with the total subsidy through a weighing coefficient σ in the government's objective.

Constraints (2)-(4) ensure geographic coverage in the aggregate sense. That is, a sufficient portion (denoted by δ) of demand nodes must be under the coverage of at least one TC within certain coverage-requirement-related distance d_0, which is assume to be independent of any specific demand node. Generally speaking, in the common practice of policy development, the government would always request some overall coverage threshold for the entire catchment area to ensure some notion of patient safety net, which has no direct relationship with the cause of an srUT error. Constraint (5) restricts the government's subsidy to be positive integer, which does not cause the loss of generality in the practical sense, since we can ensure integrality on the decision variable by changing the monetary unit of the government's subsidy. Meanwhile, this constraint can ease the algorithm design as it prevents the problem from being ill-conditioned (see the discussion in Moore and Bard (1990)).

In the lower-level model, the hospital group determines the locations for both TCs and NTCs, denoted by y, and the assignment of each demand node to TC, denoted by x. The hospital group's objective function, $Hos(\tilde{x}, \tilde{y})$, contains three parts: (a) the cost corresponding to the NTC upgrading decisions, (b) the savings corresponding to the TC downgrading decisions, and (c) the operational revenue based on the assignment decisions. We specify the upgraded and downgraded facilities with $(1 - H_j)\tilde{y}_j$ and $H_j(1 - \tilde{y}_j)$, respectively, where H_j is a binary indicator of TC vs. NTC at each location j in the original network. Constraints (7)-(8) ensure each demand node to be assigned to only one TC within a demand-node-related distance $\bar{d}_i$. Generally, $\bar{d}_i$ denotes the threshold radius of a circular region from demand node i. That is, the hospital group would prefer to ensure sufficient coverage to all potential patients for some sort of minimum level of service. Note that this differs from the government at the upper level for that this coverage requirement at the lower level is demand node specific, whereas the government is usually only concerned with some safety performance index over the entire catchment area. Constraint (9) restricts the upgrading decisions by the total amount of government subsidy. Constraint (10) specifies an additional restriction on the assignment between demand nodes and TC locations, termed the *minimum workload assignment* (MWA) rule. This rule guarantees that a TC cannot be established at location

j unless its total demand quantity exceeds a required minimum workload level, denoted by R_{min}. The MWA rule can lead to a set of facility location and demand allocation decisions such that TCs would be operated to realize sufficient use and avoid a waste of advanced medical resource.

4. Solution Method

The BIP problem (1)-(11) is intrinsically hard to solve, both theoretically and computationally. First, it is difficult to characterize the bilevel feasible region, since the lower-level optimality condition (i.e., $(x,y) \in \text{argmax}_{\tilde{x},\tilde{y}} \{Hos(\tilde{x},\tilde{y}) : (7)-(11)\}$) must be satisfied. Second, for practical values of I and J (e.g., in some U.S. states, there are typically more than 100 TCs and NTCs combined and more than 1000 zip-code areas), the lower-level model is a large-scale NP-hard problem. Third, when the subsidy s_c becomes smaller, the bilevel feasible set may get smaller, larger, unchanged, partially altered, or totally altered. Therefore, a simple iterative solution approach would not work. To address the above difficulties, we propose an exact branch-and-bound algorithm suitable to solve real-world BIP instances in a reasonable amount of time. We consider a relaxation problem of the BIP problem, which is an integer programming problem containing both upper-level and lower-level constraints but removing the lower-level optimality constraint. The relaxation problem is often referred to as the high point problem in the bilevel programming literature. It is clear that the feasible region of the relaxation problem contains the bilevel feasible region. We thus search for bilevel feasible solutions in the feasible region of the relaxation problem (i.e., our search space). In the algorithm, we iteratively solve the relaxation problem of some node problem and remove its optimal solution if it is deemed bilevel infeasible (i.e., it violates the lower-level optimality condition). Such a procedure is a standard approach in the literature. While algorithmic procedures based on high point problem are common, the novelty of our algorithm lies in the design of a branching idea with which more bilevel infeasible solutions can be removed from the search space. For our problem, we only need to focus on the situations of optimality and infeasibility because the BIP problem (1)-(11) and all the node problems are bounded. In the following, we first introduce the necessary notation and definitions. Then we will present our exact branch-and-bound algorithm. All proofs are provided in Appendix III.

4.1. Notation and definitions

We define $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$ as the set that includes all real numbers as well as positive and negative infinity. For a given set of parameters $(l,u,w \in \overline{\mathbb{R}})$, we define the node problem $\mathcal{B}(l,u,w)$ as the following parametric BIP problem:

$$\min_{s_c,x,y,z} \; Gov(s_c,y),$$

$$\begin{aligned} \text{s.t.}\ \ & (2)-(5), \\ & l \le s_c \le u, \\ & Hos(x,y) \ge w, \\ & (x,y) \in \operatorname{argmax}_{\tilde{x},\tilde{y}}\{Hos(\tilde{x},\tilde{y}) : (7)-(11)\}. \end{aligned}$$

In $\mathcal{B}(l,u,w)$, the subsidy is restricted by an interval $[l,u]$ and the net profit of the hospital group is restricted by a lower bound w. Since $\sum_{j\in J} S_j(1-H_j)\tilde{y}_j \le \sum_{j\in J} S_j$ and minimizing the subsidy is one of the government's objective, $\sum_{j\in J} S_j$ can serve as an upper bound of the subsidy. Meanwhile, $-\sum_{j\in J}|C_j^U - S_j|$ can serve as a conservative lower bound of $Hos(x,y)$, which corresponds to the situation with only negative return. Therefore, the BIP problem (1)-(11) is equivalent to $\mathcal{B}(0, \sum_{j\in J} S_j, -\sum_{j\in J}|C_j^U - S_j|)$, which will be set as the root node.

For a given set of parameters $(l,u,w \in \overline{\mathbb{R}})$, we define a relaxation problem $\mathcal{R}(l,u,w)$ as the following parametric integer programming problem:

$$\begin{aligned} \min_{s_c,x,y,z}\ \ & Gov(s_c,y), \\ \text{s.t.}\ \ & (2)-(5),(7)-(11), \\ & l \le s_c \le u, \\ & Hos(x,y) \ge w. \end{aligned}$$

Notice that variables x, y here are decision variables of the government. $\mathcal{R}(l,u,w)$ depicts the situation of the government being the sole decision-maker. In this situation, the government makes all relevant decisions (i.e., s_c, x, y) to achieve the objective, as long as it ensures the net profit of the hospital group to exceed w. It is clear that the objective of the government in this situation is better than that in $\mathcal{B}(l,u,w)$. $\mathcal{R}(l,u,w)$ is referred to as the high point problem, which was first used in Bialas and Karwan (1984) for bilevel linear programming problem and then in Moore and Bard (1990) for BMILP; however, our definition is different from theirs. In fact, $\mathcal{R}(l,u,w)$ relaxes the requirement on (x,y) so that (x,y) can be any feasible solution rather than an optimal solution to the lower-level model; as such $\mathcal{R}(l,u,w)$ provides a lower bound on $\mathcal{B}(l,u,w)$. For any value of (l,u,w), both iterative node problem $\mathcal{B}(l,u,w)$ and its relaxation problem $\mathcal{R}(l,u,w)$ are bounded, because the decision variables are all bounded in a finite set.

4.2. Bounding

Given a node problem $\mathcal{B}(\hat{l},\hat{u},\hat{w})$, we solve its relaxation problem $\mathcal{R}(\hat{l},\hat{u},\hat{w})$ and denote (s_c^R, x^R, y^R) as an optimal solution, which provides a lower bound on $\mathcal{B}(\hat{l},\hat{u},\hat{w})$. If (x^R,y^R) is optimal to $\mathcal{L}(s_c^R)$ (i.e., the lower-level model), then (s_c^R, x^R, y^R) is a bilevel feasible solution and provides an upper bound on $\mathcal{B}(\hat{l},\hat{u},\hat{w})$. Subsequently, we can characterize (s_c^R, x^R, y^R) in association with the optimality of $\mathcal{B}(\hat{l},\hat{u},\hat{w})$ as follows.

LEMMA 1. *Let* (s_c^R, x^R, y^R) *be an optimal solution to* $\mathcal{R}(\hat{l}, \hat{u}, \hat{w})$. *Then* (s_c^R, x^R, y^R) *is optimal to* $\mathcal{B}(\hat{l}, \hat{u}, \hat{w})$ *if and only if* (x^R, y^R) *is optimal to* $\mathcal{L}(s_c^R)$.

Lemma 1 is a well-known result which has been proved and used many times in the literature on discrete bilevel programming, e.g., Xu and Wang (2014), Caramia and Mari (2015), Wang and Xu (2017) and Liu et al. (2021).

In $\mathcal{R}(\hat{l}, \hat{u}, \hat{w})$, the government becomes the sole decision-maker and makes decision (s_c^R, x^R, y^R) optimizing his own objective. Thus, in $\mathcal{B}(\hat{l}, \hat{u}, \hat{w})$, the government could never obtain a better objective as the hospital group would somewhat affect the government in the negative way for her own profit maximization. If (s_c^R, x^R, y^R) happens to achieve the hospital group's optimal objective, we can conclude that (s_c^R, x^R, y^R) is bilevel optimal. Nevertheless, it is more common that the hospital group is not "satisfied" with (x^R, y^R). That is, the optimal solution to $\mathcal{L}(s_c^R)$ (denoted as (x^L, y^L)) is strictly better than (x^R, y^R), i.e., $Hos(x^L, y^L) > Hos(x^R, y^R)$. Then (s_c^R, x^R, y^R) is a bilevel infeasible solution. We thus proceed with branching to remove the bilevel infeasible solution.

4.3. Branching

In this section, we introduce a branching idea that can be used to speed up the standard branch-and-bound algorithm for BIP problem. Based on the branching idea, we present a branching rule for cases where (s_c^R, x^R, y^R) is optimal to $\mathcal{R}(\hat{l}, \hat{u}, \hat{w})$ but bilevel infeasible. The branching idea in this section tailors the well-performed enhanced branching idea proposed by Liu et al. (2021), but overcomes the previous shortage of being suboptimal if there are multiple optimal solutions in the lower-level problem.

A branching rule: *Let* (s_c^R, x^R, y^R) *be an optimal solution to* $\mathcal{R}(\hat{l}, \hat{u}, \hat{w})$. *Suppose* (x^L, y^L) *is an optimal solution to* $\mathcal{L}(s_c^R)$ *but* (x^R, y^R) *is not. The following two new node problems, denoted as* $\mathcal{B}(l^1, u^1, w^1)$ *and* $\mathcal{B}(l^2, u^2, w^2)$, *can be created from its parent node problem* $\mathcal{B}(\hat{l}, \hat{u}, \hat{w})$ *as:*

$$l^1 = \hat{l}, \quad u^1 = \sum_{j \in J} S_j(1 - H_j) y_j^L - 1, \quad w^1 = \hat{w};$$
$$l^2 = s_c^R + 1, \quad u^2 = \hat{u}, \quad w^2 = Hos(x^L, y^L).$$

To prove the validity of the branching rule, we first present Lemma 2 as follows, which constructs a subspace $\mathcal{P}$ including (s_c^R, x^R, y^R) but no bilevel feasible solutions. We remove $\mathcal{P}$ from the feasible region of $\mathcal{R}(\hat{l}, \hat{u}, \hat{w})$ so as to exclude the bilevel infeasible solution but excluding no bilevel feasible solutions. We then divide the rest of the search space into two subspaces for the branching.

LEMMA 2. *Let* (s_c^R, x^R, y^R) *be optimal to* $\mathcal{R}(\hat{l}, \hat{u}, \hat{w})$ *but bilevel infeasible. Suppose that* (x^L, y^L) *is an optimal solution to* $\mathcal{L}(s_c^R)$. *Then the following subspace:*

$$\mathcal{P} = \left\{ (s_c, x, y) : s_c \geq \sum_{j \in J} S_j(1 - H_j) y_j^L, \, Hos(x, y) < Hos(x^L, y^L) \right\}$$

contains (s_c^R, x^R, y^R) *but no bilevel feasible solutions.*

Lemma 2 here uses Lemma 3 in Xu and Wang (2014) for reference. Lemma 2 shows that if the government provides a subsidy value equal to or larger than $\sum_{j\in J} S_j(1-H_j)y_j^L$, and the hospital group achieves a profit strictly smaller than $Hos(x^L, y^L)$, then such decision cannot satisfy the hospital group. Therefore, to achieve a bilevel feasible solution, we have two choices:

(a) The government provides a subsidy value strictly smaller than $\sum_{j\in J} S_j(1-H_j)y_j^L$;

(b) The government provides a subsidy value equal to or larger than $\sum_{j\in J} S_j(1-H_j)y_j^L$ and the hospital group achieves a profit equal to or larger than $Hos(x^L, y^L)$.

The two choices correspond to the following subspaces $\mathcal{P}_1$ (choice (a)) and $\mathcal{P}_2$ (choice (b)), which are achieved by removing $\mathcal{P}$ from the search space:

$$\mathcal{P}_1 = \left\{(s_c, x, y) : s_c < \sum_{j\in J} S_j(1-H_j)y_j^L\right\},$$
$$\mathcal{P}_2 = \left\{(s_c, x, y) : s_c \geq \sum_{j\in J} S_j(1-H_j)y_j^L,\, Hos(x, y) \geq Hos(x^L, y^L)\right\}.$$

Next, we present a branching idea to further reduce the search space. Without loss of generality, we assume that S_j is integer for all $j \in J$ in the lower-level model. We present an analysis of the lower-level model in Lemma 3 as follows.

LEMMA 3. *Let* (x^L, y^L) *be an optimal solution to* $\mathcal{L}(s_c^R)$ *for some* $s_c^R \in \mathbb{Z}^+$. *If* $\sum_{j\in J} S_j(1-H_j)y_j^L < s_c^R$, *then* (x^L, y^L) *is an optimal solution to* $\mathcal{L}(s_c)$ *for any* $s_c \in \left[\sum_{j\in J} S_j(1-H_j)y_j^L, s_c^R\right]$.

Lemma 3 reminds us to search for bilevel feasible solutions within the set

$$\mathcal{S}_0(s_c^R, x^L, y^L) = \left\{(s_c, x, y) : s_c \in \left[\sum_{j\in J} S_j(1-H_j)y_j^L, s_c^R\right],\, Hos(x, y) \geq Hos(x^L, y^L),\, (7)-(11)\right\}.$$

In particular, for any $(\hat{s}_c, \hat{x}, \hat{y}) \in \mathcal{S}_0(s_c^R, x^L, y^L)$, $(\hat{x}, \hat{y})$ is feasible to $\mathcal{L}(\hat{s}_c)$ and $Hos(\hat{x}, \hat{y}) \geq Hos(x^L, y^L)$. Based on Lemma 3, (x^L, y^L) is optimal to $\mathcal{L}(\hat{s}_c)$, then $(\hat{x}, \hat{y})$ is optimal to $\mathcal{L}(\hat{s}_c)$. We select those $(\hat{s}_c, \hat{x}, \hat{y})$ from $\mathcal{S}_0(s_c^R, x^L, y^L)$ such that $(\hat{s}_c, \hat{x}, \hat{y})$ satisfies the constraints in the upper-level model so as to be bilevel feasible, i.e., $(\hat{s}_c, \hat{x}, \hat{y}) \in \mathcal{S}_1(s_c^R, x^L, y^L)$, where

$$\mathcal{S}_1(s_c^R, x^L, y^L) = \left\{(s_c, x, y) : (s_c, x, y) \in \mathcal{S}_0(s_c^R, x^L, y^L), (2)-(5)\right\}.$$

Finally, we choose the element from $\mathcal{S}_1(s_c^R, x^L, y^L)$ that obtains the optimal objective value for the upper-level model as the optimal solution to $\mathcal{B}(\hat{l}, \hat{u}, \hat{w})$ with $s_c \in \left[\sum_{j\in J} S_j(1-H_j)y_j^L, s_c^R\right]$. We sum up the idea above in Lemma 4 as follows.

LEMMA 4. *Let (x^L, y^L) be an optimal solution to $\mathcal{L}(s_c^R)$ for some s_c^R. If the following problem, denoted as $\mathcal{Q}(s_c^R, x^L, y^L)$:*

$$\begin{aligned} \min_{s_c,x,y,z} \quad & Gov(s_c, y), \\ s.t. \quad & (2)-(5), (7)-(11), \\ & Hos(x,y) \geq Hos(x^L, y^L), \\ & s_c \leq s_c^R, \\ & s_c \geq \sum_{j\in J} S_j(1-H_j)y_j^L, \end{aligned}$$

is optimal, denote the optimal solution as (s_c^Q, x^Q, y^Q), then (s_c^Q, x^Q, y^Q) is an optimal solution to $\mathcal{B}(\hat{l}, \hat{u}, \hat{w})$ with $s_c \in \left[\sum_{j\in J} S_j(1-H_j)y_j^L, s_c^R\right]$; Otherwise, there is no bilevel feasible solution to $\mathcal{B}(\hat{l}, \hat{u}, \hat{w})$ with $s_c \in \left[\sum_{j\in J} S_j(1-H_j)y_j^L, s_c^R\right]$.

Based on Lemma 4, there is no better bilevel feasible solution than (s_c^Q, x^Q, y^Q) or no bilevel feasible solution at all within the interval $s_c \in \left[\sum_{j\in J} S_j(1-H_j)y_j^L, s_c^R\right]$. Therefore, after removing the subspace $\mathcal{P}$ from the search space, we solve $\mathcal{Q}(s_c^R, x^L, y^L)$ and record the solution (if there is any), and then we carve out the set $\left[\sum_{j\in J} S_j(1-H_j)y_j^L, s_c^R\right]$ from the subspaces $\mathcal{P}_1$ and $\mathcal{P}_2$. Thus the remaining search space is reduced to the union of the following two subspaces $\mathcal{P}_1$ and $\mathcal{P}_2'$:

$$\begin{aligned} \mathcal{P}_1 &= \left\{(s_c, x, y) : s_c < \sum_{j\in J} S_j(1-H_j)y_j^L\right\}, \\ \mathcal{P}_2' &= \left\{(s_c, x, y) : s_c \geq s_c^R + 1,\, Hos(x,y) \geq Hos(x^L, y^L)\right\}. \end{aligned}$$

Note that $\mathcal{P}_1$ is unchanged while $\mathcal{P}_2'$ is a strict subset of $\mathcal{P}_2$.

The two new node problems $\mathcal{B}(l^1, u^1, w^1)$ and $\mathcal{B}(l^2, u^2, w^2)$ are created accordingly from the intersections between the feasible region of the parent node problem and the two subspaces $\mathcal{P}_1$ and $\mathcal{P}_2'$, respectively. Now we show that the two new node problems are strictly tightened as opposed to the parent node problem, by verifying (a) $u^1 < \hat{u}$; and (b) $l^2 > \hat{l}, w^2 > \hat{w}$.

For (a), we have $u^1 = \sum_{j\in J} S_j(1-H_j)y_j^L - 1 < \sum_{j\in J} S_j(1-H_j)y_j^L \leq s_c^R \leq \hat{u}$. The first inequality is obvious; the second one is valid because (x^L, y^L) is feasible to $\mathcal{L}(s_c^R)$; and the last one is valid because (s_c^R, x^R, y^R) is feasible to $\mathcal{R}(\hat{l}, \hat{u}, \hat{w})$.

For (b), we have $l^2 = s_c^R + 1 > s_c^R \geq \hat{l}$. The first inequality is obvious; the second one is valid because (s_c^R, x^R, y^R) is feasible to $\mathcal{R}(\hat{l}, \hat{u}, \hat{w})$. We also have $w^2 = Hos(x^L, y^L) > Hos(x^R, y^R) \geq \hat{w}$. The former inequality is valid by the definition of (x^L, y^L); and the latter one is valid because (s_c^R, x^R, y^R) is feasible to $\mathcal{R}(\hat{l}, \hat{u}, \hat{w})$.

Lemma 3 and Lemma 4 are suitable for the BIP with a one-dimensional variable in the upper level and upper-level variable appearing in the lower-level linear constraints. For such problem,

Lemma 3 and Lemma 4 can be embedded in a branch-and-bound framework to remove more subspace from the search space without carving out any better feasible solutions, thus improving the performance of the branch-and-bound algorithm.

4.4. Fathoming

A node problem $\mathcal{B}(\hat{l}, \hat{u}, \hat{w})$ can be fathomed in the following three ways. First, as described in section 4.2, when (s_c^R, x^R, y^R), the optimal solution to $\mathcal{R}(\hat{l}, \hat{u}, \hat{w})$, is bilevel feasible, the node problem can be fathomed. Meanwhile, (s_c^R, x^R, y^R) is saved as an incumbent solution (the best bilevel feasible solution found so far) to the BIP problem (1)-(11), along with its optimal function value ζ^R. We denote the optimal function value of the BIP problem (1)-(11) as $\zeta^* = \zeta^R$. The second way is to verify if the optimal function value of the relaxation problem is strictly smaller than ζ^*. If it is not, the node problem can be fathomed since it cannot have a feasible solution better than the incumbent. The third way is quite straightforward. If the relaxation problem has no feasible solutions, then the node problem itself must have no feasible solutions, so it can be fathomed.

We provide a detailed algorithm description in Appendix II. The finite termination of the algorithm is verified naturally by the boundness of the BIP problem. The correctness of the algorithm is stated as follows.

THEOREM 1. *Our algorithm outputs an optimal network design for our SNrDP or declares the infeasibility of the input BIP instance.*

The proofs of the above results are provided in Appendix III. Note that these results hold for BIPs with general integer variables in both levels. To justify the computational efficiency, we in Appendix IV report an additional computational study to compare our BIP algorithm with two state-of-the-art BIP solution methods (Xu and Wang 2014, Fischetti et al. 2017) on a generalized version of our subsidized network design optimization problem.

5. A Real-world Case Study

In this section, we conduct a case study on the subsidized trauma care network redesign problem for a catchment area in the midwest United State to verify the applicability of our methodology. Firstly, we describe the data on a trauma care network in a catchment area in the midwestern U.S., and present the BIP model parameter values in the real-world case. Then, we report a baseline case where sensitive parameters are set to be the mean values of the respective value ranges. Finally, we report two sets of sensitivity analysis experiments on model parameters to examine the factors that influence the network design. We implement the algorithm in Matlab and set CPLEX 12.9 as the ILP solver. We conduct the numerical experiments pertaining to this case study on a computer cluster consisting of Intel(R) Xeon(R) E5-2678v3 CPUs with 2.5 GHz and 64 GB of RAM.

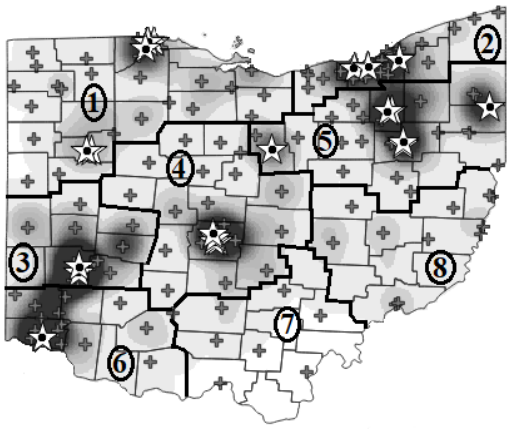


Figure 2 The heat map of trauma incidents and the locations of TCs and NTCs in the original network.

5.1. The dataset

We obtain a sample of over 7,000 deidentified trauma incidents that take place in the catchment area in the midwestern U.S. in 2012, including the location (i.e., longitude and latitude) and the field assessment (i.e., Injury Severity Score). After removing missing data, we are left with 6,002 records, which is about 10% of trauma injury incidents occurred in this area in 2012. We generate the geographic distribution of the sample incidents and confirm that the sample distribution is consistent with that of overall incidents.

The trauma care network of the area includes 21 TCs and 140 NTCs. Then 161 care facility locations ($J = \{1, ..., 161\}$) are geocoded in terms of their latitudes and longitudes. Then the original network configuration (i.e., H_j's) is specified. Figure 2 illustrates the heat map of 6,002 incidents and the location of TCs (labeled by stars) and NTCs (labeled by crosses), which is divided by the Department of Health into eight regions. Figure 2 shows that TCs were clustered in region 1, 2, 3, 4 and 5 (mostly metropolitan areas), whereas the vast relatively rural area in south and southeast essentially has no TCs (i.e., regions 7 and 8). To estimate the coverage requirement of both the government and the hospital group, we use the Haversine formula to estimate the radius of the eight regions (i.e., $\bar{d}_i$). We set d_0 as the minimum value of $\{\bar{d}_1, ..., \bar{d}_8\}$. The required coverage level of the government is set as $\delta = 0.8$, which implies an average coverage area of d_0 achieving a level of 80%.

From the dataset, we acquire the number of trauma incidents at each demand node (T_i). We extract the values of financial parameters ($C^T, C^{NT}, C^U_j, C^D_j, R^T_{ij}$) based on the OHSU (Oregon Health and Science University) study and the literature (Newgard et al. 2013, Zocchi et al. 2016). We identify [\$500,000, \$1,000,000] to be the value range for annualized monetary compensation for upgrading each NTC (S_j), based on the inputs of multiple experts (Singer 2012). Due to lack of

specific numbers, we assume the compensation does not differ among candidate locations. For the MWA rule considered, we identify [40, 65] to be the value range for the minimum workload R_{min} (Jansen et al. 2018). In the baseline case, we set S_j to be \$750,000 and R_{min} to be 50.

There are two model parameters in the government's problem, namely γ and σ, which play a vital role in determining the network design. Note that γ quantifies the monetary loss due to the health hazard of each srUT error. Considering that the trauma care literature often suggests under-triage as a more critical concern than over-triage to the government, we set γ to be 10,000 in the baseline case, which is four times as much as the coefficient associated with over-triage (i.e., $C^T - C^{NT}$). Note that σ reflects the government's willingness to subsidize the TC upgrading in the tradeoff of system-wide spending. In the baseline case, we set the weight for subsidy as 0.1.

To model $O(y)$ and $U(y)$, we first build a high-fidelity simulator, which emulates a field-triage decision hierarchy generalized from a notional protocol suggested in the medical literature (Jansen et al. 2018, Hirpara et al. 2020). As we elect to solve the relaxation problems in our branch-and-bound algorithm by an off-the-shelf MIP solver, we then approximate $O(y)$ and $U(y)$ as close-form polynomial functions with respect to network design y. To construct the approximate functions, we apply symbolic regression coupled with stratified sampling to avoid the significant computational burden while maintaining sufficient accuracy of the approximations. Meanwhile, we ensure the resultant approximate polynomial functions of $O(y)$ and $U(y)$ to be acceptable by the MIP solver when incorporated in the objective function of the relaxation problems in our branch-and-bound algorithm, e.g., these functions are required to be quadratic functions. We provide details of the simulator and the approximation functions construction in Appendices V and VI respectively.

For any given network y, we ran the high-fidelity simulator to calculate the number of srUT errors $U(y)$ and srOT errors $O(y)$. We define the "network improvement" of y as

$$\frac{(C^T - C^{NT})(O(H) - O(y)) + \gamma(U(H) - U(y))}{(C^T - C^{NT})O(H) + \gamma U(H)} \times 100\%,$$

where H denotes the original network.

5.2. The optimal network redesign in the baseline case

Using our algorithm, we solve the real-world instance in the baseline case. Table 3 presents the details of the optimal network redesign in the baseline case, including the numbers of TCs and NTCs (labeled as "TCs/NTCs"), the numbers of NTC upgrading and TC downgrading (labeled as "up/down"), the network improvement (labeled as "N-Impro"), and the value of subsidy s_c (in million dollars). Figure 3 and Figure 4 show the heat maps of mistriage in the original network and the redesigned network, respectively. We have four observations as follows.

Table 2 The optimal network redesign in the baseline case

TCs/NTCs	up/down	N-Impro	s_c
17/144	8/12	11.14%	6

Observation 1. *In the redesigned network, some NTCs have been upgraded to TCs in regions 1, 3, 7 and 8, where the volume of srUT errors is high and TCs are absent. As a result, the srUT errors are reduced at the corresponding positions.*

Observation 2. *In the redesigned network, some TCs have been downgraded to NTCs in regions 1, 2, 3, 4 and 5, where the volume of srOT errors is high and TCs are clustered. As a result, the srOT errors are reduced at the corresponding positions.*

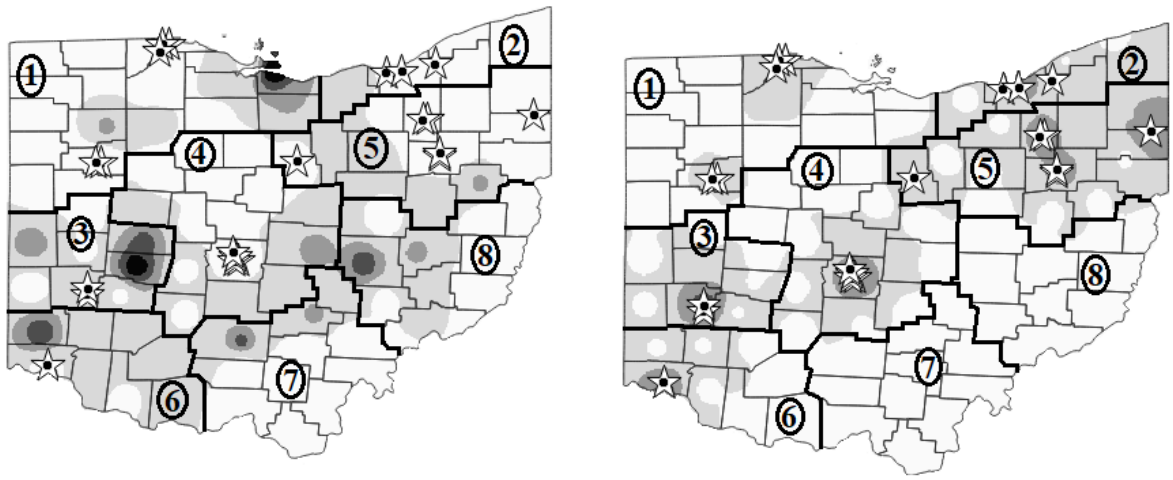
In the redesigned network, there would be 12 TC downgrading and 8 NTC upgrading, which corresponds to a modest decrease on trauma specialty care capacity. But a subsidy of 6 million dollars would be needed to incentivize the geographic redistribution of trauma care facilities. It suggests that the main issue to this area is not a lack of trauma specialty care capacity, rather it is about finding ways to better distribute TCs.

Observation 3. *The upgrading of NTCs may result in an increase of srOT errors in the surrounding areas. For example, the upgrading of three NTCs in region 8 has resulted in an increase of srOT errors at the corresponding position.*

Observation 4. *The downgrading of TCs may result in an increase of srUT errors in the surrounding areas. For example, the downgrading of two TCs in region 5 has resulted in an increase of srUT errors at the corresponding position.*

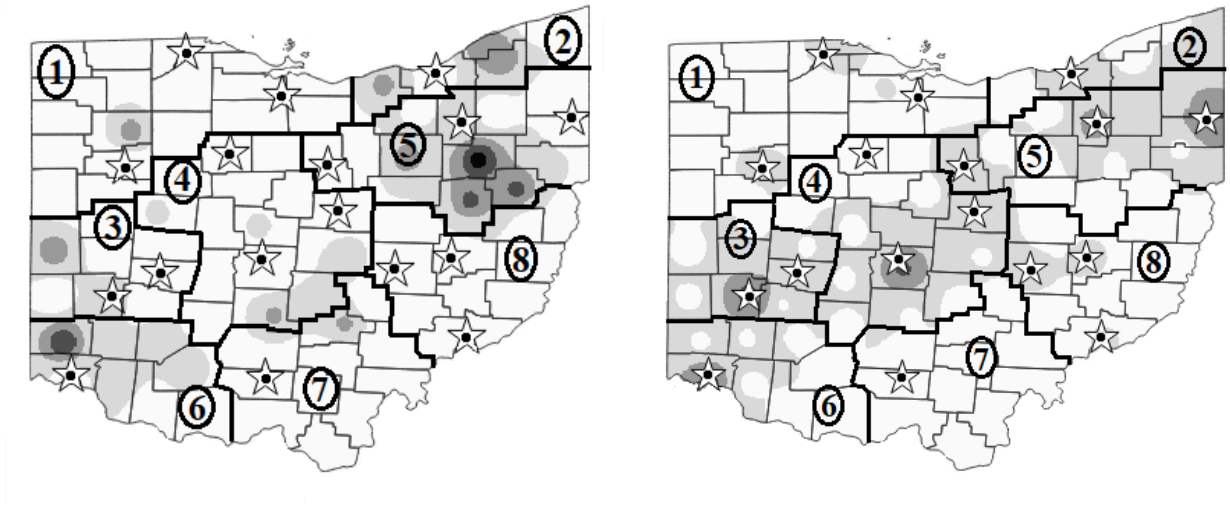
There is a tradeoff between NTC upgrading and TC downgrading, since NTC upgrading (TC downgrading) may result in an increase of srOT (srUT) errors. The optimal upgrading/downgrading decisions are made based on a consideration between the negative effect of srOT errors and srUT errors. Overall, in the optimal network redesign, both srUT and srOT errors would decrease, associated with a "network improvement" value of 11.14%.

To point out the differences between decision-making under the bilevel framework and the relaxation problem (i.e., the situation where the government is the sole decision-maker), we calculate the optimal values achieved by the government and the hospital group. The government can achieve an objective value of about 1.200×10^7 in the relaxation problem and 1.335×10^7 in the bilevel program. The hospital group can achieve a profit value of about 2.300×10^7 in the relaxation problem and 3.025×10^7 in the bilevel program. That is, the government's objective value is up by more than 11% when it is the sole decision-maker; whereas the hospital group's objective value is up by more than 30% when its voice also must be heard.



srUT errors: 182 srOT errors: 2559

Figure 3 Heat maps of mistriage (srUT on the left and srOT on the right) in the orginal network.



srUT errors: 182 → 138 srOT errors: 2559 → 2388

Figure 4 Heat maps of mistriage (srUT on the left and srOT on the right) in the redesigned network.

5.3. Sensitivity analysis

We then conduct sensitivity analysis on γ/σ and R_{min}/S_j to assess how the identified network redesign and associated metrics would be affected. For γ, we consider values 5000, 7500, 10000, 12500, and 15000, which equals two, three, four, five, and six times of the weight assigned to over-triage errors. Larger γ values imply increased preference of the government towards srUT error reduction. For σ, we consider values 0.01, 0.1, 1.0. Larger σ values imply the government is less willing to spend on the subsidization. Table 4 reports the optimal solutions. In each cell, the five

values from top to bottom report the number of srUT errors, the number of srOT errors, the network improvement value, the number of TCs, and the subsidy amount (in million dollars).

Table 3 Sensitivity analysis results based on γ/σ

σ \ γ	5000	7500	10000	12500	15000
0.01	158	134	134	116	109
	2283	2399	2399	2546	2580
	11.14%	10.36%	11.38%	11.21%	13.06%
	15	19	19	18	20
	4.5	7.5	7.5	5.25	7.5
0.1	158	138	138	116	109
	2283	2388	2388	2546	2580
	11.14%	10.25%	11.14%	11.21%	13.06%
	15	17	17	18	20
	4.5	6	6	5.25	7.5
1.0	158	153	153	153	143
	2283	2333	2333	2333	2339
	11.14%	10.28%	10.64%	10.96%	12.95%
	15	15	15	15	16
	4.5	4.5	4.5	4.5	5.25

With the sensitivity analysis on γ/σ, we have the following observations.

(a) When γ increases, the number of srUT errors decreases whereas the number of srOT errors increases. This observation confirms the intuition that the network redesign would reduce the number of srUT errors as the value of γ increases (i.e., the harm of srUT is regarded more critical).

(b) When σ decreases, the number of TCs, number of srOT errors, network improvement value, and subsidy amount all increase, whereas the number of srUT errors decreases. For example, when $\gamma = 15,000$, as σ decreases from 1.0 to 0.1, the number of srOT errors increases by 241 while the number of srUT errors decreases by 34, which results in network performance increase from 12.95% to 13.06%. This suggests that if the government is more willing to subsidize (small σ), the corresponding optimal network redesign would achieve larger relative improvement over the baseline. In other words, subsidization can be a key measure to improve the network performance.

For R_{min}, we consider values 0, 50, 100 and 150. For S_j, we consider values 0.5, 0.75 and 1 million dollars. Table 5 reports the optimal solutions: the numbers of UT and OT errors, the network improvement value (labeled as "N-Impro"), the number of TCs, and the subsidy amount s_c (in million dollars).

With the sensitivity analysis experiments based on R_{min}/S_j, we have the following observations.

(a) When R_{min} increases from 0 (which corresponds to no minimum workload concern), the number of TCs decreases and the network improvement value increases. When R_{min} continues to increase, the results no longer change, until there is no solution when R_{min} reaches 150.

Table 4 Sensitivity analysis results based on R_{min}/S_j

Parameter		UT	OT	N-Impro	TC	s_c
R_{min}	0	116	2546	8.43%	18	5.25
	50	138	2388	11.14%	17	6
	100	138	2388	11.14%	17	6
	150	–	–	–	–	–
S_j	0.5	128	2514	7.94%	16	3
	0.75	138	2388	11.14%	17	6
	1.0	138	2388	11.14%	17	8

(b) When S_j increases from 0.5 million dollars to 1 million dollars, the number of TCs, subsidy amount, and network improvement value all increase.

Our results based on R_{min} suggest that it is critical to set this value appropriately as it has substantial impact on the network improvement. The results based on S_j suggest that an increase on the requested monetary compensation can help improve the network performance.

6. Conclusions

In this paper, we study a subsidized trauma care network redesign problem considering two decision-makers, the government and the hospital group. We aim to identify a promising network redesign plan that includes the amount of government subsidy, the location of TCs/NTCs, and the assignment of demand nodes to TCs. Our goal is to reduce system-wide health hazard due to under-triage errors and unnecessary care cost due to over-triage errors, without requiring a large amount of government subsidy for the network redesign.

We employ bilevel integer programming to model the subsidized network design optimization problem. Discrete bilevel programming models are challenging to solve, especially for large-scale instances. In this paper, we leverage a novel branching idea that can exclude additional infeasible solutions and suboptimal solutions at each iteration. Generally speaking, our branching idea can be embedded into any standard branch-and-bound framework to speed up the algorithm. Compared with a well-performing exact solution algorithm in the literature, we verify the efficiency of our branching idea with randomly generated instances. Using our methodology, we conduct a case study based on real trauma incident data and geography information. Our case study calls for a more scattering pattern of locating TCs in this area, by which one can see an overall improvement of around 11%. More encouragingly, such improvement is especially worthwhile when the tension is largely due to poor TC access of rural residents and pronounced health risks associated with it.

Our work is of significant relevance to public policy. Care for trauma injuries and other types of sudden injuries is of undeniable criticality to regional/district governments worldwide, including department of public safety and agency of family and social administration. Nevertheless, under

a market economy, the government is not a care provider, but a potentially care financier. Hence, it is important from a policy viewpoint to wisely address the public-private tension, which is key in general to spatial redistribution of trauma specialty care resource. In this work, we investigate the viability of government subsidization to facilitating care network redesign and to ensuring system-wide improvement in both reducing health risks and curbing care spending.

We can extend the subsidized discrete location optimization model studied in this paper to many public sector OR research areas, and address the need of effectively establishing public-private partnership for aligning location-specific multi-class service demands within a location-specific multi-class service network. This may be especially useful when the common impression is that the overall system-wide service capacity is insufficient. Instead of increasing the overall capacity blindly, our methodology can lead to a socioeconomically beneficial solution as it can help improve the geographic distribution of the capacity such as appropriate services are more likely delivered at the right time to the right place.

Our future research can be undertaken in the following directions. First, it could be more accurate to discard the approximation functions for estimating $O(y)$ and $U(y)$. Instead, we will incorporate pricing in the bilevel optimization framework based on our high-fidelity simulator so as to evaluate and generate promising network designs "on the fly". Second, there exists an alternative setting where the government is allowed to determine a subsidy "menu" for NTC upgrading rather than requesting S_j, the asking price on upgrading each NTC in the hospital group, in advance. The resultant bilevel programming is much more difficult to solve for the appearing nonlinear constraints linking the two levels, i.e., in constraints $\sum_{j\in J} S_j(1-H_j)\tilde{y}_j \le s_c$, S_j's are upper-level decision variables and $\tilde{y}_j$'s are lower-level decision variables. Third, in a more resource deprived catchment area, specialized trauma care need should be assessed periodically and some form of resource rationing may be needed through allocating the resource to different locations. Thus, it would be interesting to combing the decision making of these two aspects in this tiered network: location-specific capacity planning and facility location analysis. Fourth, there can be multiple hospital groups that have independent authorities and competing interests on network redesign. Therefore, we plan to study the subsidized network redesign problem with two or more decision-makers at the lower level for future research. Finally, conducting more real case studies based on other areas (e.g., Indiana, USA) and with consideration of additional important policy concerns (e.g., geographic fairness) is an interesting topic for future research. Now, some of us are actively collaborating with the Division of Trauma System/Injury Prevention Program at the Indiana State Department of Health and their scientific advisory board.

Acknowledgments

References

Ahmadi-Javid, A., Seyedi, P., Syam, S.S. (2017). A survey of healthcare facility location. *Computers and Operations Research.* 44:223-263.

Aksen, D., Aras, N. (2012). A bilevel fixed charge location model for facilities under imminent attack. *Computers and Operations Research.* 39(7):1364-1381.

Aksen D., Aras N., Karaarslan A.G. (2009). Design and analysis of government subsidized collection systems for incentive-dependent returns. *International Journal of Production Economics.* 119(2):308-327.

Aksen, D., Akca, S.Ş., Aras, N. (2014). A bilevel partial interdiction problem with capacitated facilities and demand outsourcing. *Computers and Operations Research.* 41:346-358.

ATS (2016). Trauma center levels explained. *American Trauma Society.* Available at `http://www.amtrauma.org/?page=traumalevels`.

Avraamidou, S., Pistikopoulos, E.N. (2019). A Multi-parametric optimization approach for bilevel mixed-integer linear and quadratic programming problems. *Computers and Chemical Engineering.* 125:98-113.

Baxt, W.G., Jones, G., Fortlage, D. (1990). The trauma triage rule: A new, resource-based approach to the prehospital identification of major trauma victims. *Annals of Emergency Medicine.* 19(12): 1401-1406.

Bélanger, V., Lanzarone, E., Nicoletta, V., Ruiz, A., Soriano, P. (2020). A recursive simulation-optimization framework for the ambulance location and dispatching problem. *European Journal of Operational Research.* 286(2): 713-725.

Bhadury, J., Eiselt, H.A. (2012). Optimizing subsidies for the location of distribution centers. *Annals of Regional Science.* 48(1): 247-261.

Bialas, W.F., Karwan, M.H. (1984). Two-level linear programming. *Management science.* 30 (8): 1004–1020.

Branas, C.C., MacKenzie, E.J., Williams, J.C., Schwab, C.W., Teter, H.M., Flanigan, M.C., Blatt, A.J., ReVelle, C.S. (2005). Access to trauma centers in the United States. *JAMA.* 293(21): 2626-2633.

Branas, C.C., Wolff, C.S., Williams, J., Margolis, G., Carr, B.G. (2013). Simulating changes to emergency care resources to compare system effectiveness. *Journal of Clinical Epidemiology.* 66(8): 57-64.

Branas, C.C., MacKenzie, E.J., ReVelle, C.S. (2000). A trauma resource allocation model for ambulances and hospitals. *Health Services Research.* 35(2): 489.

Branas, C.C., Revelle, C.S. (2001). An iterative switching heuristic to locate hospitals and helicopters. *Socio-Economic Planning Sciences.* 35(1): 11-30.

Brotcorne, L., Hanafi, S., Mansi, R. (2013). One-level reformulation of the bilevel knapsack problem using dynamic programming. *Discrete Optimization.* 10(1): 1-10.

Brown, J.B., Rosengart, M.R., Billiar, T.R., Peitzman, A.B., Sperry, J.L. (2016). Geographic distribution of trauma centers and injury related mortality in the United States. *The Journal of Trauma and Acute Care Surgery.* 80(1): 42.

Carr, B.G., Branas, C.C. (2010). Traumamaps.org - Trauma center maps. *University of Pennsylvania Cartographic Modeling Laboratory.*

Carr, B.G., Walsh, L., Williams, J.C., Pryor, J.P., Branas, C.C. (2016). A geographic simulation model for the treatment of trauma patients in disasters. *Prehospital and Disaster Medicine.* 31(4): 413-421.

Caramia, M., Mari, R. (2015). Enhanced exact algorithms for discrete bilevel linear problems. *Optimization Letters.* 9(7): 1447-1468.

Caramia, M., Mari, R. (2016). A decomposition approach to solve a bilevel capacitated facility location problem with equity constraints. *Optimization Letters.* 10(5): 997-1019.

Chang, D. (2016). New trauma center to open at Jackson South following state approval. *Miami Herald.* Available at `http://www.miamiherald.com/news/health-care/article75303107.html`.

Cho, S.H., Jang, H., Lee, T., Turner, J. (2014). Simultaneous location of trauma centers and helicopters for emergency medical Service planning. *Operations Research.* 62(4): 751-771.

Côté, M.J., Syam, S.S., Vogel, W.B., Cowper, D.C. (2007). A mixed integer programming model to locate traumatic brain injury treatment units in the Department of Veterans Affairs: A case study. *Health Care Management Science.* 10(3): 253-267.

Daskin, M.S., Dean, L.K. (2004). Location of health care facilities. *Operations Research and Health Care.* Springer, Boston, MA. 43-76.

Dempe, S., Kalashnikov, V., Pérez-Valdés, G.A., Kalashnykova, N.I. (2011). Natural gas bilevel cash-out problem: Convergence of a penalty function method. *European Journal of Operational Research.* 215(3): 532-538.

Dempe, S., Kalashnikov, V., RíOs-Mercado, R.Z. (2005). Discrete bilevel programming: Application to a natural gas cash-out problem. *European Journal of Operational Research.* 166(2):469-488.

DeNegre, S.T., Ralphs, T.K. (2009). A branch-and-cut algorithm for integer bilevel linear programs. *Operations Research and Cyber-Infrastructure.* Springer, Boston, MA. 65-78.

Erdemir, E.T., Batta, R., Rogerson, P.A., Blatt, A., Flanigan, M. (2010). Joint ground and air emergency medical services coverage models: A greedy heuristic solution approach. *European Journal of Operational Research.* 207(2):736-749.

Enayati, S., Mayorga, M. E., Rajagopalan, H. K., Saydam, C. (2018). Real-time ambulance redeployment approach to improve service coverage with fair and restricted workload for EMS providers. *Omega.* 79:67-80.

Fáisca, N.P., Dua, V., Rustem, B., Saraiva, P.M., Pistikopoulos, E.N. (2007). Parametric global optimisation for bilevel programming. *Journal of Global Optimization.* 38(4):609-623.

Fischetti, M., Ljubić, I., Monaci, M., Sinnl, M. (2019). Interdiction games and monotonicity, with application to knapsack problems. *INFORMS Journal on Computing.* 31(2): 390-410.

Fischetti, M., Ljubić, I., Monaci, M., Sinnl, M. (2018). On the use of intersection cuts for bilevel optimization. *Mathematical Programming.* 172(1-2):77-103.

Fischetti, M., Ljubić, I., Monaci, M., Sinnl, M.. (2017). A new general-purpose algorithm for mixed-integer bilevel linear programs. *Operations Research.* 65(6):1615-1637.

Fischetti, M., Ljubić, I., Monaci, M., Sinnl, M. (2016). Intersection cuts for bilevel optimization. *International Conference on Integer Programming and Combinatorial Optimization.* Springer, Cham. 77-88.

Galewitz, P. (2012). Trauma centers springing up as profits rise. *USA Today.* Available at `http://usatoday30.usatoday.com/money/business/story/2012/09/24/trauma-centers-springing-up-as-profits-rise/57838766/1`.

Ghaffarinasab, N., Atayi, R. (2018). An implicit enumeration algorithm for the hub interdiction median problem with fortification. *European Journal of Operational Research.* 267(1):23-39.

Ghaffarinasab, N., Motallebzadeh, A. (2018). Hub interdiction problem variants: Models and metaheuristic solution algorithms. *European Journal of Operational Research.* 267(2):496-512.

Gutjahr, W.J., Dzubur, N. (2016). Bi-objective bilevel optimization of distribution center locations considering user equilibria. *Transportation Research Part E: Logistics and Transportation Review.* 85:1-22.

Griffin, P.M., Scherrer, C.R., Swann, J.L. (2008). Optimization of community health center locations and service offerings with statistical need estimation. *IIE Transactions.* 40(9):880-892.

Harper, P. R., Shahani, A. K., Gallagher, J. E., Bowie, C. (2005). Planning health services with explicit geographical considerations: a stochastic location–allocation approach. *Omega.* 33(2):141-152.

Haywood, A. B., Lunday, B. J., Robbins, M. J. (2022). Intruder detection and interdiction modeling: A bilevel programming approach for ballistic missile defense asset location. *Omega.* 110: 102640.

Hemmati, M., Smith, J.C. (2016). A mixed-integer bilevel programming approach for a competitive prioritized set covering problem. *Discrete Optimization.* 20:105-134.

Vaishnav, M., Hirpara, S., Parikh, P.J., Kong, N., Parikh, P. (2020), Locating trauma centers considering patient safety. *Working paper.*

Jansen, J.O., Morrison, J.J., Wang, H., He, S., Lawrenson, R., Hutchison, J.D., Campbell, M.K. (2015). Access to specialist care: Optimizing the geographic configuration of trauma systems. *Journal of Trauma and Acute Care Surgery.* 79(5): 756.

Jansen, J.O., Morrison, J.J., Wang, H., Lawrenson, R.M., Egan, G., He, S., Campbell, M.K. (2014). Optimizing trauma system design: the GEOS (Geospatial Evaluation of Systems of Trauma Care) approach. *Journal of Trauma and Acute Care Surgery.* 76(4):1035-1040.

Jansen, J.O., Moore, E.E., Wang, H., Morrison, J.J., Hutchison, J.D., Campbell, M.K., Sauaia, A. (2018). Maximizing geographical efficiency: An analysis of the configuration of Colorado's trauma system. *Journal of Trauma and Acute Care Surgery.* 84(5):762-770.

Jeroslow, R. G. (1985). The polynomial hierarchy and a simple model for competitive analysis. *Mathematical Programming.* 32(2):146-164.

Jia, H., Ordóñez, F., Dessouky, M. (2007). A modeling framework for facility location of medical services for large-scale emergencies. *IIE Transactions.* 39(1):41-55.

Jones, C.M.C., Cushman, J.T., Lerner, E.B., Fisher, S., Seplaki, C.L., Veazie, P.J., Wasserman, E.B., Dozier, A., Shah, M.N. (2016). Prehospital trauma triage decision-making: A model of what happens between the 9-1-1 call and the hospital. *Prehospital Emergency Care.* 20(1):6-14.

Kalashnikov, V.V., Ríos-Mercado, R.Z. (2006). A natural gas cash-out problem: A bilevel programming framework and a penalty function method. *Optimization and Engineering.* 7(4):403-420.

Keçici, S., Aras, N., Verter, V. (2012). Incorporating the threat of terrorist attacks in the design of public service facility networks. *Optimization Letters.* 6(6):1101-1121.

Knight, V. A., Harper, P. R., Smith, L. (2012). Ambulance allocation for maximal survival with heterogeneous outcome measures. *Omega.* 40(6): 918-926.

Küçükaydin, H., Aras, N., Altınel, I.K. (2011). Competitive facility location problem with attractiveness adjustment of the follower: A bilevel programming model and its solution. *European Journal of Operational Research.* 208(3):206-220.

Lampariello, L., Sagratella, S. (2017). A bridge between bilevel programs and Nash games. *Journal of Optimization Theory and Applications.* 174(2): 613-635.

Lavlinskii, S. M., Panin, A. A., Plyasunov, A. V. (2021). Bilevel Models for Socially Oriented Strategic Planning in the Natural Resources Sector. *International Conference on Mathematical Optimization Theory and Operations Research.* Springer, Cham.

Lavlinskii, S. M., Panin, A. A., Plyasunov, A. V. (2018). Public-private partnership models with tax incentives: numerical analysis of solutions. *International Conference on Optimization Problems and Their Applications.* Springer, Cham, 220-234.

Lavlinskii, S. M., Panin, A. A., Plyasunov, A. V. (2015). A bilevel planning model for public–private partnership. *Automation and remote control* 76(11): 1976-1987.

Lee, T., Jang, H. (2018). An iterative method for simultaneously locating trauma centers and helicopters through the planning horizon. *Operations Research for Health Care.* 19:185-196.

Li, X., Zhao, Z., Zhu, X., Wyatt, T. (2011). Covering models and optimization techniques for emergency response facility location and planning: A review. *Mathematical Methods of Operations Research.* 74(3):281-310.

Liberatore, F., Scaparra, M.P., Daskin, M.S. (2012). Hedging against disruptions with ripple effects in location analysis. *Omega.* 40(1):21-30.

Liu, S., Wang, M., Kong, N., Hu, X. (2021). An enhanced branch-and-bound algorithm for bilevel integer linear programming. *European Journal of Operational Research.* 291: 661-679.

MacKenzie, E.J., Rivara, F.P., Jurkovich, G.J., Nathens, A.B., Frey, K.P., Egleston, B.L., Salkever, D.S., Scharfstein, D.O. (2006). A national evaluation of the effect of trauma-center care on mortality. *New England Journal of Medicine.* 354(4):366-378.

Mackersie, R.C. (2006). History of trauma field triage development and the American College of Surgeons criteria. *Prehospital Emergency Care.* 10(3):287-294.

Mestre, A.M., Oliveira, M.D., Barbosa-Póvoa, A.P. (2015). Location-allocation approaches for hospital network planning under uncertainty. *European Journal of Operational Research.* 240(3):791-806.

Moore, J.T., Bard, J.F. (1990). The mixed integer linear bilevel programming problem. *Operations research.* 38(5):911-921.

Newgard, C.D., Staudenmayer, K., Hsia, R.Y., Mann, N.C., Bulger, E.M., Holmes, J.F., Fleischman, R., Gorman, K., Haukoos, J., McConnell, K.J. (2013). The cost of overtriage: More than one-third of low-risk injured patients were taken to major trauma centers. *Health Affairs.* 32(9):1591-1599.

Newgard, C.D., Zive, D., Holmes, J.F., Bulger, E.M., Staudenmayer, K., Liao, M., Rea, T., Hsia, R.Y., Wang, N.E., Fleischman, R., Jui, J., Mann, N.C., Haukoos, J.S., Sporer, K.A., Gubler, K.D., Hedges, J.R. (2011). A multisite assessment of the American College of Surgeons Committee on trauma field triage decision scheme for identifying seriously injured children and adults. *Journal of the American College of Surgeons.* 213(6):709-721.

NTI (2016). Trauma Statistics. *National Trauma Institute.* Available at `http://www.nationaltraumainstitute.org/home/traumastatistics.html`.

Parikh, P.P., Parikh, P., Guthrie, B., Mamer, L., Whitmill, M., Erskine, T.,Woods, R., Saxe, J. (2017). Impact of triage guidelines on prehospital triage: comparison of guidelines with a statistical model. *Journal of Surgical Research.* 220:255-260.

Parikh, P.P., Parikh, P., Mamer, L., McCarthy, M.C., Sakran, J.V. (2019). Association of System-Level Factors With Secondary Overtriage in Trauma Patients. *JAMA Surgery.* 154(1):19-25.

Rahman, S.U., Smith, D.K. (2000). Use of Location-Allocation Models in Health Service Development Planning in Developing Nations. *European Journal of Operational Research.* 123(3):437-452.

Rhee, P., Joseph, B., Pandit, V., Aziz, H., Vercruysse, G., Kulvatunyou, N., Friese, R.S. (2014). Increasing trauma deaths in the United States. *Annals of Surgery.* 260(1):13-21.

Rodríguez González, S. (2020). Addressing the principal-agent problem in public private partnerships via mixed-integer bi-level linear programming. *Master's thesis, Uniandes.*

Saharidis, G.K., Ierapetritou, M.G. (2009). Resolution method for mixed integer bi-level linear problems based on decomposition technique. *Journal of Global Optimization.* 44(1):29-51.

Sasser, S.M., Hunt, R.C., Faul, M., Sugerman, D., Pearson, W.S., Dulski, T., Wald, M.M., Jurkovich, G.J., Newgard, C.D., Lerner, E.B., Cooper, A., Wang, S.C., Henry, M.C., Salomone, J.P., Galli, R.L. (2011). Guidelines for field triage of injured patients: recommendations of the National Expert Panel on Field Triage. *Morbidity and Mortality Weekly Report: Recommendations and Reports.* 61(1):1-20.

Scaparra, M.P., Church, R.L. (2008). A bilevel mixed-integer program for critical infrastructure protection planning. *Computers and Operations Research.* 35(6):1905-1923.

Singer, S. (2012). Tenet wants to triple subsidy for trauma care; HCA offers to do it for free. *Palm Beach Post.* Available at `https://www.palmbeachpost.com/business/tenet-wants-triple-subsidy-for-trauma-care-hca-offers-for-free/b8NgdOHfOdRpw9jEKdnHeM/`.

Syam, S.S., Côté, M.J. (2010). A location-allocation model for service providers with application to not-for-profit health care organizations. *Omega.* 38(3-4):157-166.

Sudtachat, K., Mayorga, M. E., Mclay, L. A. (2016). A nested-compliance table policy for emergency medical service systems under relocation. *Omega.* 58: 154-168.

Tahernejad, S., Ralphs, T.K., DeNegre, S.T. (2020). A branch-and-cut algorithm for mixed integer bilevel linear optimization problems and its implementation. *Mathematical Programming Computation.* 1-40.

Tang, Y., Richard, J.P.P., Smith, J.C. (2016). A class of algorithms for mixed-integer bilevel min-max optimization. *Journal of Global Optimization.* 66(2): 225-262.

Tanınmış, K., Aras, N., Altınel, İ. K. (2022). Improved $x-$space algorithm for min-max bilevel problems with an application to misinformation spread in social networks. *European Journal of Operational Research.* 297(1): 40-52.

Verter, V., Lapierre, S.D. (2002). Location of preventive health care facilities. *Annals of Operations Research.* 110(1-4): 123-132.

Vidyarthi, N., Jayaswal, S. (2014). Efficient solution of a class of location-allocation problems with stochastic demand and congestion. *Computers and Operations Research.* 48: 20-30.

Wang, L., Xu, P. (2017). The watermelon algorithm for the bilevel integer linear programming problem. *SIAM Journal on Optimization.* 27(3): 1403-1430.

Xu, P., Wang, L. (2014). An exact algorithm for the bilevel mixed integer linear programming problem under three simplifying assumptions. *Computers and Operations Research.* 41: 309-318.

Yue, D., Gao, J., Zeng, B., You, F. (2019). A projection-based reformulation and decomposition algorithm for global optimization of a class of mixed integer bilevel linear programs. *Journal of Global Optimization.* 73(1):27-57.

Zare, M.H., Borrero, J.S., Zeng, B., Prokopyev, O.A. (2019). A note on linearized reformulations for a class of bilevel linear integer problems. *Annals of Operations Research.* 272(1-2):99-117.

Zeng, B., An, Y. (2014). Solving bilevel mixed integer program by reformulations and decomposition. *Optimization Online.* 1-34.

Zhang, J., Özaltın, O.Y. (2017). A branch-and-cut algorithm for discrete bilevel linear programs. *Optimization Online.*

Zhang, Y., Berman, O., Marcotte, P., Verter, V. (2010). A bilevel model for preventive healthcare facility network design with congestion. *IIE Transactions.* 42(12):865-880.

Zocchi, M.S., Hsia, R.Y., Carr, B.G., Sarani, B., Pines, J.M. (2016). Comparison of mortality and costs at trauma and nontrauma centers for minor and moderately severe injuries in California. *Annals of emergency medicine.* 67(1):56-67.

Appendix I. Real-World Scenarios

An 86-year-old male was involved in a motor vehicle crash in a rural part of southeast. Given that he could not open his eyes, exhibited no verbal communication, and had an altered level of consciousness, the EMS personnel determined his injuries severe enough to require immediate trauma care. Because the closest trauma center was nearly an hour away, he was transported in an ambulance to the nearest community hospital (non-trauma center). Although he was eventually transferred from this hospital to that trauma center a day later, due to the significant delay in specialty trauma care (i.e., comprehensive care for the full spectrum of injuries beyond the initial assessment and resuscitation phase), he succumbed to his injuries.

In another scenario, a 30-year-old female fell down during an early morning jog in the downtown of a metropolitan city. Although her injuries were not severe, considering the cluster of trauma centers in the vicinity, she was transported to one of them (10 minutes away), instead of an appropriate non-trauma center (only 15 minutes away). She was discharged home the very next day from this hospital. A few days later, she received a medical bill of over $24,000, out of which nearly $11,000 was associated with the fixed cost of visiting that trauma center.

So could the male patient's life be saved if a trauma center was located much closer to the scene? Could the female patient be sent to a non-trauma center just a few more minutes away and not be slapped with a huge medical bill instead? Could the time spent on her be better utilized by the trauma staff towards caring for other sicker trauma patients at that hospital? Such questions motivated our research on optimizing the trauma care network design.

Appendix II. Algorithm Description

We present our exact branch-and-bound algorithm, which outputs a global optimal solution $(s_c^*, x^*, y^*, \zeta^*)$ to the BIP problem (1)-(11) as follows. $(s_c^* = \emptyset, x^* = \emptyset, y^* = \emptyset, \zeta^* = +\infty)$ denotes the infeasible case. z records the relaxation problem's objective value for bounding purpose and $\mathcal{N}$ denotes the set of active nodes in the branch-and-bound tree.

Algorithm 1

Step 0 (Initialization): Create the root node $\mathcal{B}(l^0, u^0, w^0)$, which is parameterized by $l^0 = 0, u^0 = \sum_{j\in J} S_j, w^0 = -\sum_{j\in J} |C_j^U - S_j|$. Initialize $s_c^* = \emptyset, x^* = \emptyset, y^* = \emptyset, \zeta^* = +\infty, \mathcal{N} = \{\mathcal{B}(l^0, u^0, w^0)\}$, and $z^0 = -\infty$. Go to Step 1.

Step 1 (Node management): For any node problem $\mathcal{B}(l^k, u^k, w^k) \in \mathcal{N}$ such that $z^k \geq \zeta^*$ or $l^k > u^k$, remove node $\mathcal{B}(l^k, u^k, w^k)$ from $\mathcal{N}$.

if $\mathcal{N} = \emptyset$ **then**

if $s_c^* \neq \emptyset$ **then**

1(a) return $(s_c^*, x^*, y^*, \zeta^*)$ is an optimal solution to the BIP problem (1)-(11).

else

1(b) return the BIP problem (1)-(11) is infeasible.

end if

else

1(c) select a node $\mathcal{B}(l^k, u^k, w^k)$ from $\mathcal{N}$, set $(\hat{l} = l^k, \hat{u} = u^k, \hat{w} = w^k)$, remove the node from $\mathcal{N}$. Go to Step 2.

end if

Step 2 (Relaxation): Solve $\mathcal{R}(\hat{l}, \hat{u}, \hat{w})$.

if $\mathcal{R}(\hat{l}, \hat{u}, \hat{w})$ is infeasible **then**

2(a) go to Step 1.

else

let (s_c^R, x^R, y^R) denote an optimal solution to $\mathcal{R}(\hat{l}, \hat{u}, \hat{w})$.

if $Gov(s_c^R, y^R) \geq \zeta^*$ **then**

2(b) go to Step 1.

else

2(c) go to Step 3.

end if

end if

Step 3 (Lower level): Solve $\mathcal{L}(s_c^R)$, and denote the optimal solution as (x^L, y^L)

if $Hos(x^R, y^R) = Hos(x^L, y^L)$ **then**

3(a) update $(s_c^* = s_c^R, x^* = x^R, y^* = y^R, \zeta^* = Gov(s_c^R, y^R))$, and go to Step 1.

else

if $\sum_{i\in I, j\in J} C_{ij} y_j^L \geq \delta |I|$ and $Gov(s_c^R, x^L, y^L) < \zeta^*$ **then**

update $s_c^* = s_c^R, x^* = x^L, y^* = y^L, \zeta^* = Gov(s_c^R, y^L)$

if $Gov(s_c^R, y^R) \geq \zeta^*$ **then**

3(b) go to Step 1.

end if

end if

if $s_c^R > \sum_{j\in J} S_j(1 - H_j) y_j^L$, $\mathcal{Q}(s_c^R, x^L, y^L)$ is optimal (denote the optimal solution as (s_c^Q, x^Q, y^Q)), and $Gov(s_c^Q, y^Q) < \zeta^*$ **then**

update $s_c^* = s_c^Q, x^* = x^Q, y^* = y^Q, \zeta^* = Gov(s_c^Q, y^Q)$

if $Gov(s_c^R, y^R) \geq \zeta^*$ **then**

3(c) go to Step 1.

end if

end if

3(d) go to Step 4.

end if

Step 4 (Branching): Create two new nodes based on the branching rule, set their value of z to be $Gov(s_c^R, y^R)$, and add the two new nodes to $\mathcal{N}$, and go to Step 1.

Appendix III. Proofs for the Algorithm

PROOF of LEMMA 1. The "only if" direction is a direct result of the bilevel optimality of (s_c^R, x^R, y^R). For the "if" direction, (x^R, y^R) being optimal to $\mathcal{L}(s_c^R)$ implies that (s_c^R, x^R, y^R) is a bilevel feasible solution to $\mathcal{B}(\hat{l}, \hat{u}, \hat{w})$, and thus it provides an upper bound on $\mathcal{B}(\hat{l}, \hat{u}, \hat{w})$. Meanwhile, (s_c^R, x^R, y^R) achieves a lower bound on $\mathcal{B}(\hat{l}, \hat{u}, \hat{w})$ since (s_c^R, x^R, y^R) is optimal to $\mathcal{R}(\hat{l}, \hat{u}, \hat{w})$. Therefore, the "if" statement follows. The proof of Lemma 1 is completed. ■

PROOF of LEMMA 2. Since (x^L, y^L) is an optimal solution to $\mathcal{L}(s_c^R)$ but (x^R, y^R) is not, we have $\sum_{j\in J} S_j(1-H_j)y_j^L \le s_c^R$ and $Hos(x^R, y^R) < Hos(x^L, y^L)$, that is $(s_c^R, x^R, y^R) \in \mathcal{P}$. For any $(\bar{s}_c, \bar{x}, \bar{y}) \in \mathcal{P}$, we show that $(\bar{x}, \bar{y})$ cannot be optimal to $\mathcal{L}(\bar{s}_c)$. Actually, we have $\sum_{j\in J} S_j(1-H_j)y_j^L \le \bar{s}_c$ and $Hos(\bar{x}, \bar{y}) < Hos(x^L, y^L)$. This implies that (x^L, y^L) is a feasible solution to $\mathcal{L}(\bar{s}_c)$ and superior to $(\bar{x}, \bar{y})$. The proof of Lemma 2 is completed. ■

PROOF of LEMMA 3. If there exists some $s_c^0 \in \left[\sum_{j\in J} S_j(1-H_j)y_j^L, s_c^R\right]$ such that (x^L, y^L) is not optimal to $\mathcal{L}(s_c^0)$, then there exists some (x^0, y^0) satisfying constraints (5)-(9) and $Hos(x^0, y^0) > Hos(x^L, y^L)$. Then we have that $\sum_{j\in J} S_j(1-H_j)y_j^L \le s_c^0 \le s_c^R$, which implies that (x^0, y^0) is also feasible to $\mathcal{L}(s_c^R)$ and $Hos(x^0, y^0) > Hos(x^L, y^L)$. This violates the optimality of (x^L, y^L) for $\mathcal{L}(s_c^R)$. The proof of Lemma 3 is completed. ■

PROOF of LEMMA 4. If (s_c^Q, x^Q, y^Q) is optimal to $\mathcal{Q}(s_c^R, x^L, y^L)$, then $s_c^Q \in \left[\sum_{j\in J} S_j(1-H_j)y_j^L, s_c^R\right]$ and $Hos(x^Q, y^Q) \ge Hos(x^L, y^L)$. Following Lemma 3, we have (x^Q, y^Q) is optimal to $\mathcal{L}(s_c^Q)$. Since (s_c^Q, x^Q, y^Q) satisfies (2)-(3), (s_c^Q, x^Q, y^Q) is bilevel feasible. Further, (s_c^Q, x^Q, y^Q) is bilevel optimal to $\mathcal{B}(\hat{l}, \hat{u}, \hat{w})$ with $s_c \in \left[\sum_{j\in J} S_j(1-H_j)y_j^L, s_c^R\right]$ since (s_c^Q, x^Q, y^Q) is optimal to $\mathcal{Q}(s_c^R, x^L, y^L)$. The proof of Lemma 4 is completed. ■

PROOF of THEOREM 1. We prove the theorem by showing the correctness at all decision points in the algorithm.

Steps 1(a), 1(b), 1(c), 2(a), 2(b), 2(c), 3(a) and 3(d) are standard procedures in a branch-and-bound framework.

Step 3(b) shows that if $s_c = \sum_{j\in J} S_j(1-H_j)y_j^L$, we only need to test if (s_c^R, x^L, y^L) is bilevel feasible and superior rather than solving $\mathcal{Q}(s_c^R, x^L, y^L)$ before branching.

Step 3(c) determines that (s_c^Q, x^Q, y^Q) is a superior solution based on Lemma 4.

Step 4 carves out a piece of the (s_c, x, y) space from the feasible region of the current node and creates two new subproblems based on the branching rule.

The proof of Theorem 1 is completed. ■

Appendix IV. Algorithm Description and Computational Results for a Generalized BIP Formulation

To systematically investigate the performance of our algorithm, we generalize the BIP formulation for the subsidized network redesign problem and adapt our exact branch-and-bound algorithm to it in this appendix. We report a numerical study that compares our algorithm with two state-of-the-art general-purpose BIP solution methods (Xu and Wang 2014, Fischetti et al. 2017).

This generalized BIP formulation is such that the upper-level problem contains only one integer decision variable; the upper-level decision variable should only be involved in one lower-level linear constraint; and both the upper- and lower-level variables are bounded. The generalized BIP formulation is thus presented as follows:

$$\begin{aligned}
\min_{x,y}\ & F(x,y),\\
\text{s.t.}\ & A_1x+B_1y\le b_1,\\
& 0\le x\le X,\\
& x\in\mathbb{Z},\\
& y\in\arg\max_{\tilde{y}}\{f(\tilde{y}): A_2x+B_2\tilde{y}\le b_2, B_3\tilde{y}\le b_3, 0\le\tilde{y}\le Y, \tilde{y}\in\mathbb{Z}^{n_2}\},
\end{aligned}$$

where $A_1, b_1\in\mathbb{R}^{m_1\times 1}, B_1\in\mathbb{R}^{m_1\times n_2}, X\in\mathbb{Z}_+$, $A_2, b_2\in\mathbb{Z}$, $B_2\in\mathbb{Z}^{1\times n_2}$, $B_3\in\mathbb{R}^{m_2\times n_2}$, $b_3\in\mathbb{R}^{m_2\times 1}$, $Y\in\mathbb{Z}_+^{n_2\times 1}$.

The lemmas and theories proposed in Section 4 can be extended to the generalized BIP formulation. In the following paragraph, we provide the formulation of $\mathcal{B}(\hat{l},\hat{u},\hat{w})$, $\mathcal{R}(\hat{l},\hat{u},\hat{w})$, $\mathcal{L}(x^R)$, $\mathcal{Q}(x^R,y^L)$, and the related " branching rule" for the generalized BIP formulation, since they are crucial for the implementation of the algorithm.

The node problem $\mathcal{B}(l,u,w)$ can be formulated as follows:

$$\begin{aligned}
\min_{x,y}\ & F(x,y),\\
\text{s.t.}\ & A_1x+B_1y\le b_1,\\
& 0\le x\le X,\\
& x\in\mathbb{Z},\\
& y\in\arg\max_{\tilde{y}}\{f(\tilde{y}): A_2x+B_2\tilde{y}\le b_2, B_3\tilde{y}\le b_3, 0\le\tilde{y}\le Y, \tilde{y}\in\mathbb{Z}^{n_2}\},\\
& l\le A_2x\le u,\\
& f(y)\ge w.
\end{aligned}$$

The relaxation problem $\mathcal{R}(l,u,w)$ can be formulated as follows:

$$\begin{aligned}
\min_{x,y}\ & F(x,y),\\
\text{s.t.}\ & A_1x+B_1y\le b_1,\\
& 0\le x\le X,\\
& x\in\mathbb{Z},\\
& A_2x+B_2y\le b_2,\\
& B_3y\le b_3,\\
& 0\le y\le Y,\\
& y\in\mathbb{Z}^{n_2},\\
& l\le A_2x\le u,\\
& f(y)\ge w.
\end{aligned}$$

The lower lever problem $\mathcal{L}(x^R)$ can be formulated as follows:

$$\begin{aligned}
\max_{y}\ & f(y),\\
\text{s.t.}\ & A_2x^R+B_2y\le b_2,\\
& B_3y\le b_3,\\
& 0\le y\le Y,\\
& y\in\mathbb{Z}^{n_2}.
\end{aligned}$$

The problem $\mathcal{Q}(x^R,y^L)$ can be formulated as follows:

$$\begin{aligned}
\min_{x,y}\ & F(x,y),\\
\text{s.t.}\ & A_1x+B_1y\le b_1,\\
& 0\le x\le X,\\
& x\in\mathbb{Z},\\
& A_2x+B_2y\le b_2,\\
& B_3y\le b_3,\\
& 0\le y\le Y,\\
& y\in\mathbb{Z}^{n_2},\\
& f(y)\ge f(y^L),\\
& A_2x^R\le A_2x\le b_2-B_2y^L.
\end{aligned}$$

A branching rule: *Let* (x^R, y^R) *be an optimal solution to* $\mathcal{R}(\hat{l}, \hat{u}, \hat{w})$*. Suppose* y^L *is an optimal solution to* $\mathcal{L}(x^R)$ *but* y^R *is not. The following two new node problems, denoted as* $\mathcal{B}(l^1, u^1, w^1)$ *and* $\mathcal{B}(l^2, u^2, w^2)$*, can be created from its parent node problem* $\mathcal{B}(\hat{l}, \hat{u}, \hat{w})$ *as:*

$$l^1 = b_2 - B_2 y^L + 1, \quad u^1 = \hat{u}, \quad w^1 = \hat{w};$$

$$l^2 = \hat{l}, \quad u^2 = A_2 x^R - 1, \quad w^2 = f(y^L).$$

We present our exact branch-and-bound algorithm for the generalized BIP formulation (denoted as "the BIP problem" in the algorithm) as follows. The algorithm outputs a global optimal solution (x^*, y^*, ζ^*). $(x^* = \emptyset, y^* = \emptyset, \zeta^* = +\infty)$ denotes the infeasible case. z records the relaxation problem's objective value for bounding purpose and $\mathcal{N}$ denotes the set of active nodes in the branch-and-bound tree.

Algorithm 2

Step 0 (Initialization): Create the root node $\mathcal{B}(l^0, u^0, w^0)$, which is parameterized by $l^0 = \min\{0, A_2 X\}, u^0 = \max\{0, A_2 X\}, w^0 = -\infty$. Initialize $x^* = \emptyset, y^* = \emptyset, \zeta^* = +\infty, \mathcal{N} = \{\mathcal{B}(l^0, u^0, w^0)\}$, and $z^0 = -\infty$. Go to Step 1.

Step 1 (Node management): For any node problem $\mathcal{B}(l^k, u^k, w^k) \in \mathcal{N}$ such that $z^k \geq \zeta^*$ or $l^k > u^k$, remove node $\mathcal{B}(l^k, u^k, w^k)$ from $\mathcal{N}$.

if $\mathcal{N} = \emptyset$ **then**

  if $x^* \neq \emptyset$ **then**

    1(a) return (x^*, y^*, ζ^*) is an optimal solution to the BIP problem.

  else

    1(b) return the BIP problem is infeasible.

  end if

else

  1(c) select a node $\mathcal{B}(l^k, u^k, w^k)$ from $\mathcal{N}$, set $(\hat{l} = l^k, \hat{u} = u^k, \hat{w} = w^k)$, remove the node from $\mathcal{N}$. Go to Step 2.

end if

Step 2 (Relaxation): Solve $\mathcal{R}(\hat{l}, \hat{u}, \hat{w})$.
  if $\mathcal{R}(\hat{l}, \hat{u}, \hat{w})$ is infeasible **then**
    2(a) go to Step 1.
  else
    let (x^R, y^R) denote an optimal solution to $\mathcal{R}(\hat{l}, \hat{u}, \hat{w})$.
    if $F(x^R, y^R) \geq \zeta^*$ **then**
      2(b) go to Step 1.
    else
      2(c) go to Step 3.
    end if
  end if

Step 3 (Lower level): Solve $\mathcal{L}(x^R)$, and denote the optimal solution as y^L
  if $f(y^R) = f(y^L)$ **then**
    3(a) update $(x^* = x^R, y^* = y^R, \zeta^* = F(x^R, y^R))$, and go to Step 1.
  else
    if $A_1 x^R + B_1 y^L \leq b_1$ and $F(x^R, y^L) < \zeta^*$ **then**
      update $x^* = x^R, y^* = y^L, \zeta^* = F(x^R, y^L)$
      if $F(x^R, y^R) \geq \zeta^*$ **then**
        3(b) go to Step 1.
      end if
    end if
    if $A_2 x^R + b_2 y^L < b_2, \mathcal{Q}(x^R, y^L)$ is optimal (denote the optimal solution as (x^Q, y^Q)), and $F(x^Q, y^Q) < \zeta^*$ **then**
      update $x^* = x^Q, y^* = y^Q, \zeta^* = F(x^Q, y^Q)$
      if $F(x^R, y^R) \geq \zeta^*$ **then**
        3(c) go to Step 1.
      end if
    end if
    3(d) go to Step 4.
  end if

Step 4 (Branching): Create two new nodes based on "A branching rule", set their value of z to be $F(x^R, y^R)$, and add the two new nodes to $\mathcal{N}$, and go to Step 1.

We implemented the algorithm in Matlab and set CPLEX 12.9 as the ILP solver. The computational experiments were conducted on a desktop computer with 2.29GHz CPU and 8 GB of RAM. In view of a tractable BIP relaxation problem, we considered linear functions for the objective functions in the upper and the lower level as follows:

$$F(x, y) = c^T x + d_1^T y,$$
$$f(y) = d_2^T y,$$

where $c \in \mathbb{R}$, $d_1, d_2 \in \mathbb{R}^{n_2 \times 1}$. We created 7 sets of instances with $n_1 = m_1 = 1$, $n_2 \in \{5000, 6000, ..., 10000, 20000\}$, and $m_2 = 0.1n_2$. For each (n_2, m_2) pair, we randomly generated 10 instances. Thus there were 70 instances in total. The upper-level and the lower-level elements are all integers or real numbers which are uniformly distributed within a certain range: c, d_1 and d_2 are within the set $[-50, 50]$; b_1 is within the set $[30, 130]$; b_2 and b_3 are within the set [10,110]; A_1, B_1, A_2, B_2, and B_3 are within the set $[0, 10]$. x and y have bounds set to be $X = 1000$ and $Y = 1000^{n_2}$, respectively.

We solved the instances by three methods. The first method, labeled as MIX++, is based on the algorithm proposed by Fischetti et al. (2017), which utilizes intersection cuts. The second method, labeled as Alg^{B}, implements the branching rule proposed by Xu and Wang (2014). Finally, our method, labeled as Alg^{E}, employs the enhanced branching rule. In the literature, MIX++ has shown the best performance among tested instances. For MIX++, we used a solver coded and made publicly available by the authors in Fischetti et al. (2017). We self coded other two methods, which differ by the branching ideas (i.e., Alg^{B} follows Lemma 2 and Lemma 3, and Alg^{E} follows Lemmas 2-4 as an enhancement). Table EC.1 reports the average computation times (in seconds) of MIX++, Alg^{B} and Alg^{E} for each set of 10 instances. "–" denotes that MIX++ cannot solve the instances of size $n_2 = 10000$ and $n_2 = 20000$ within 20000 wall-clock seconds. In the last column, the average computation times of Alg^{B} and Alg^{E} for all 70 instances are reported. Our results show that Alg^{E} can outperform Alg^{B} by 27% to 56% on the average computation time among the instance sets, and by over 51% on average over all the 70 instances. On the other hand, the results show that MIX++ is not suitable to solve the BIP instances of the type we study in this paper. The advantage of Alg^{E} is more noticeable as the size of the lower-level model (n_2) increases.

The 70 instances created in this section as well as our solutions were posted online at https://engineering.purdue.edu/KongLab/research/BILPInstances/.

Table EC.1 Average computation time (in seconds)

n_2	5000	6000	7000	8000	9000	10000	20000	Average
MIX++	5326	7697	12141	11135	15326	–	–	–
Alg^{B}	96	216	762	1002	1735	2020	10019	2264
Alg^{E}	70	147	475	605	914	1094	4414	1103

Appendix V. A High-Fidelity Simulator

A high-fidelity simulator for system-related under-triage and over-triage errors was developed by one of the authors and his colleagues. It is based on a notional field triage protocol, adapted from that in Jansen et al. (2018). The protocol captures the hierarchical decision-making process that the EMS providers are recommended to use during field triage.

To model the EMS decisions based on transport times, we introduced two threshold values: (i) "access" threshold (time based) for transport to a TC and (ii) "bypass" threshold (time based) for transport to an NTC. The first threshold value helps determine if a case would lead to an srUT error. That is, if the attending patient were severely injured, it would be an srUT error. The second threshold value helps determine if a case would result in an srOT error. That is, if the attending patient were non-severely injured, it would be an srOT error. Further, in line with the existing trauma literature, we used Injury Severity Score (ISS) as an index for the severity of injuries at the scene. ISS is a post-hoc metric evaluated after the patient's arrival at the hospital. Note that while the first threshold was used in Jansen et al. (2018), whereas the second one has never been discussed in the literature. In that sense, our notional protocol is more general than previous work. In addition to our review of the literature, we also made observations at a leading EMS agency in our region. Figure EC.1 shows a schematic of the notional protocol.

According to the notional protocol, for severely injured patients, the EMS staff first check if a TC is accessible within the "access" threshold. If yes, then the patient is transported to the TC. If no, then they check if an air ambulance can be called in to transport the patient to the nearest TC. However, if the sum of the inbound-to-scene, loading, and transport-to-TC is higher than the "access" threshold, then the EMS would most likely take the patient to a nearby NTC, resulting in an srUT error. On tho other hand, non-severely injured patients should be taken to an NTC, if the additional time to reach a TC is within the "bypass" threshold, then the EMS often takes the patient to the TC, resulting in an srOT error. The reasons for such an srOT error can vary; TC's reputation, the-bigger-the-hospital-the-better-the-care, patient/family choice, insurance situation, and even negotiated contracts between the EMS and TC.

This high-fidelity simulator was used in our study as a computational tool to evaluate $O(y)$ and $U(y)$ for any arbitrary network design y. When y changes, the nearest TC and NTC may change for each demand node. However, this simulator does not provide closed-form expressions of $O(y)$ and $U(y)$ with respect to y. Therefore, to solve the bilevel integer programming efficiently, we replaced them with closed-form approximations.

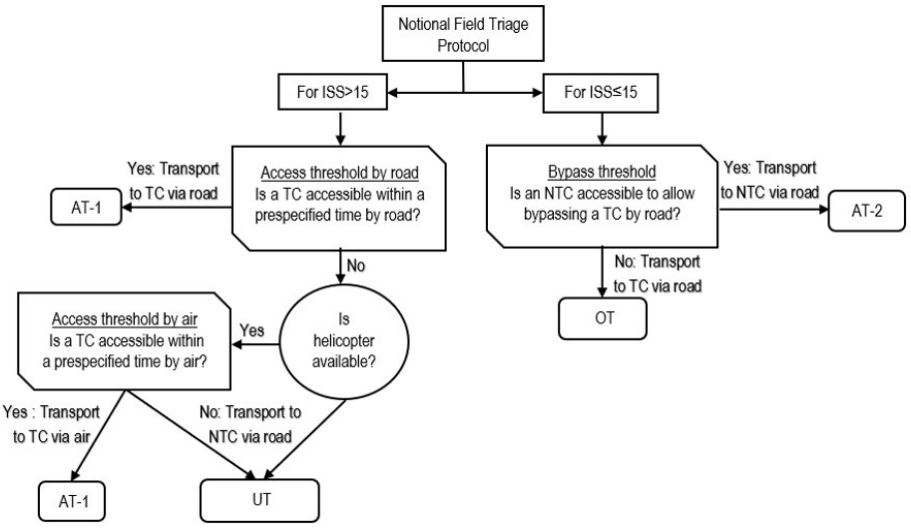


Figure EC.1 Notional field triage protocol.

Appendix VI. Approximations of $U(y)$ and $O(y)$

To efficiently solve the bilevel integer programming problem, we replaced the high-fidelity simulator of srOT and srUT errors with its respective approximations of $U(y)$ and $O(y)$. To fully quantify the mappings $y \to O(y)$ and $y \to U(y)$, we need to evaluate each y from a total of 2^{161} candidate designs (i.e., $\{(y_1, ..., y_{161})\}_{y_j \in \{0,1\}, j=1,...,161}$) based off the 161 candidate facility locations in the area. We considered that it is computationally prohibitive to perform bilevel integer optimization upon the individually based simulator over a large cohort of trauma incidents. Hence, we first resorted to a response surface approach with a carefully designed stratified sampling scheme.

We divided the whole sample space into 7 strata. In the 1st stratum, we flipped the trauma care designation (i.e., from TC to NTC; from NTC to TC) at only one location on the basis of the original network design, denoted as y^0. This resulted in 161 samples in the 1st stratum: $\{y^1, ..., y^{161}\}$. We took all the samples and computed $|U(y^k) - U(y^0)| + |O(y^k) - O(y^0)|$ for each sample y^k, $k = 1, \ldots, 161$, with $O(y)$ and $U(y)$ being outputs of the simulation. We then ranked the locations based on the above value in a descending order. We then identified 23 *one-move* redesigns on the top of the list based on some threshold. We labeled the 23 corresponding locations as significant and the other 138 locations as non-significant. We denoted y^P to be the network design with the designations of all 138 non-significant locations changed simultaneously.

In the 2nd stratum, we flipped the designations of 2 to 6 locations among the 23 significant locations based on y^0 and y^P, respectively. There are $2(C_{23}^2 + C_{23}^3 + ... + C_{23}^6)$ samples in the 2nd stratum. We used random sampling to evaluate 20 samples (sample network designs) from this stratum via the simulation.

In the 3rd stratum, we flipped the designations of 7 to 10 locations among the 23 significant locations based on y^0 and y^P, respectively. There are $2(C_{23}^7 + C_{23}^8 + ... + C_{23}^{10})$ samples in the 3rd stratum. We used random sampling to evaluate 20 samples from this stratum via the simulation.

In the 4th stratum, we flipped the destinations of 11 to 13 locations among the 23 significant locations based on y^0 and y^P, respectively. There are $2(C_{23}^{11} + C_{23}^{12} + C_{23}^{13})$ samples in the 4th stratum. We used the uniform random sampling to evaluate 20 samples from this stratum via the simulation.

In the 5th stratum, we flipped the designations of 14 to 17 locations among the 23 significant locations based on y^0 and y^P, respectively. There are $2(C_{23}^{14} + C_{23}^{15} + ... + C_{23}^{17})$ samples in the 5th stratum. We used random sampling to take 20 samples from this stratum via the simulation.

In the 6th stratum, we flipped the designations of 18 to 23 locations among the 23 significant locations based on y^0 and y^P, respectively. There are $2(C_{23}^{18} + C_{23}^{19} + ... + C_{23}^{23})$ samples in the 6th stratum. We used random sampling to take 20 samples from this stratum via the simulation.

The 7th stratum is a complementary set of the previous six strata. In this stratum, we flipped the designation of at least one but not all the locations among the 138 non-significant locations together with at least one of the 23 significant locations based on y^0. In addition, we flipped the designation of at least two of the 138 non-significant locations based on y^0. There are $(C_{138}^1 + ... + C_{138}^{137}) \cdot (C_{23}^1 + ... + C_{23}^{23}) + (C_{138}^2 + ... + C_{138}^{138})$ samples from the 7th stratum. We used random sampling to evaluate 20 samples from this stratum via the simulation.

In summary, we took a total of 402 samples along the above sampling process, which correspond to 402 samples candidate network designs y's. With the help of Eureqa (`www.nutonian.com/products/eureqa/`), a symbolic regression software package, we conducted symbolic regression to fit the two closed-form functions y vs. $O(y)$ and y vs. $U(y)$. We chose the approximations of $U(y)$ and $O(y)$ based on the output fit and size values.