Prompt and Reliable Medical Evacuation with Air Ambulances

Miguel A. Lejeune (George Washington University), François Margot (Ecole Supérieure de la Santé),
Alan Delgado de Oliveira (Microsoft)

Keywords: medical evacuation, chance constraint, endogenous uncertainty, air ambulances

1 INTRODUCTION

The timely and reliable medical evacuation (MEDEVAC) of injured soldiers on the battlefield is a primary concern for every military force and relies on determining the location of medical treatment facilities (MTF) and air ambulances and their dispatch (Shackelford *et al.*, 2024). Air ambulances increase survival rates because they are faster than ground transportation, can land at casualty collection points (CCP) not reachable by ground vehicles, and have a rescue hoist to lift casualties inside. The US Combat Casualty Care Research Program supports the "Force 2025 and Beyond" initiative that focuses upon the timeliness of MEDEVACs and their inherent uncertainty in order to increase survival rates. The **Golden Hour** military doctrine requires that every seriously injured soldier receives proper medical care within one hour. While responding in one hour is critical, it is widely argued that survival rates should be further improved with MEDEVAC plans that ensures that the time for the wounded to reach an MTF is **reliably minimal**. This study implements this approach and provides a **probabilistic guarantee** of fulfilling this goal and maximizing the survival chance and functional recovery of the wounded.

Contributions: First, we implement the response time minimization approach and propose a joint chance-constrained MEDEVAC model with exogenous and endogenous uncertainty to design a reliable and time-sensitive MEDEVAC network performing well against most possible scenarios. Endogenous uncertainty occurs when the probability distribution of random variables are altered by the decisions taken. Modeling the dependency connecting random and decision variables is critical for emergency transportation models. Erkut *et al.* (2008) argue that emergency health care models accounting for endogenous uncertainty "*not only result in better coverage estimates, but also cause coverage to be a better proxy for lives saved"*. Second, we design a reformulation and algorithmic framework for such chance-constrained models relying on the Boolean modeling approach and a new nonlinear branch and bound (BB) algorithm that solves a conic integer problem at each node. Third, we carry out numerical tests to ascertain tractability and applicability.

2 Methodology

2.1 Chance-Constrained Model with Exogenous and Endogenous Uncertainty

Description and notations: The proposed chance-constrained MEDEVAC model maximizes the expected number of severe casualties evacuated without delay with a prescribed reliability level. The model defines the location of MTFs, the deployment of air ambulances, and their dispatch to CCPs so that casualties can be evacuated as swiftly as possible. Going beyond the Golden Hour doctrine, we use an objective function that reflects the importance of response times since the survival chance of a wounded decreases with the time between injury and life-sustaining treatment. Our model incentivizes the fastest possible MEDEVAC by maximizing the number of MEDEVACs that can be immediately responded to and incorporates a chance constraint to provide a reliability guarantee on achieving this goal.

The set of possible MTF locations where air ambulances can be deployed is I and the the set of CCPs is J. The number of casualties at CCP j is $\mu_j \in J$. The binary variable z_i indicates whether an air ambulance is stationed at MTF i, $p \in (0, 1]$ is the prescribed probability level, and N is the number of deployed air ambulances. The *normalized* service time s_{ij} is the fraction of one hour needed by an air ambulance at i to carry out an evacuation at j. The parameter M_{ij} is the maximum number of evacuations from i to j in one hour. The integer variable x_{ij} is the number of requests at j serviced by an air ambulance at i. The integer variable λ_i is the number of MEDEVACs carried out within one hour with an ambulance at i. The variable y_i is the number of casualties evacuated without delay with probability p to MTF i.

Uncertainty: We model the level of available MEDEVAC resources at MTFs as an **exogenous** random variable, as the literature stresses the uncertainty surrounding MEDEVAC crews and supplies. We assume that the MEDEVAC resource level can fall below the *baseline* level by a certain percentage represented by a unitless random coefficient $\xi_i \in [0, 1], i \in I$ defining the relative decrease in available resources at MTF i. We model the underlying uncertainty with a set Ω of joint scenarios. An air ambulance is typically assigned to multiple MEDEVAC requests and can be busy when an request arrives, thereby causing delays, increasing response time, and lowering the survival chance of the wounded. We model the availability of air ambulances as an **endogenous** uncertainty using the busy probability concept, i.e., the probability that a vehicle is not immediately available. In general, the busy probabilities of vehicles can seldom be computed a priori. We model the availability of an air ambulance as a Bernoulli random variable whose expected value is defined **endogenously** and set equal to the complement of the busy probability of the ambulance. The busy probability ζ_i of ambulance i is the ambulance normalized workload which depends on location and assignment decisions, i.e. the weighted sum of the requests assigned to an air ambulance i : $\zeta_i = \sum_{j \in J} s_{ij} x_{ij}, i \in I$. Clearly, the expected availability $(1 - \zeta_i)$ of an air ambulance decreases linearly with its workload and is defined endogenously, i.e., by the assignment decisions. Finally, note that we calculate an **individual** busy probability (1b) for each air ambulance.

Formulation: The MEDEVAC model **M1** with exogenous and endogenous uncertainty

$$\mathbf{M1}: \quad \max \sum_{i \in I} y_i \tag{1a}$$

$$\text{s.to} \quad \mathbb{P}\left(\lambda_i(1-\zeta_i) - y_i(1+\xi_i) \geq 0\ ,\ i \in I\right) \geq p \tag{1b}$$

$$\mathcal{MILP} := \begin{cases} \zeta_i = \sum\limits_{j \in J} s_{ij}\, x_{ij} & i \in I \quad (1c) \\ \zeta_i \leq z_i & i \in I \quad (1d) \\ \lambda_i = \sum\limits_{j \in J} x_{ij} & i \in I \quad (1e) \\ x_{ij} \leq M_{ij}\, z_i & i \in I, j \in J \quad (1f) \\ \sum\limits_{i \in I} x_{ij} = \mu_j & j \in J \quad (1g) \\ \sum\limits_{i \in I} z_i \leq N & \quad (1h) \\ z_i \in \{0,1\}\ ,\ x_{ij} \in \mathbb{Z}_+ & i \in I \quad (1i) \\ 0 \leq x_{ij} \leq u^x_{ij}\ ,\ 0 \leq \lambda_i, y_i \leq u^\lambda_i & i \in I, j \in J \quad (1j) \end{cases}$$

is an MINLP joint chance-constrained problem and has a nonconvex continuous relaxation. The joint chance constraint (1b) includes one stochastic inequality for each air ambulance and requires each to hold jointly with probability at least equal to p. The stochastic inequalities in (1b) include exogenous as well as endogenous uncertainties and are combinatorial and nonlinear. The expression $(1 - \zeta_i)$ is the probability of an air ambulance i being immediately available and λ_i is the number of requests assigned to air ambulance i. The expression $(1 - \zeta_i)\lambda_i$ in (1b) represents the expected number of requests serviced without delay with the baseline resource level. Correspondingly, y_i is the probabilistic number of MEDEVACs serviced without delay with reliability level p: y_i decreases as the relative decrease ξ_i in available resources rises. The objective (1a) is thus the maximization of the probabilistic number of severe casualties that can be evacuated without delay. The linking constraint (1c) specifies the coupling between decision and random variables, and defines how assignment decisions x_{ij} affect air ambulance availability. As the Golden Hour doctrine requires the completion of all evacuations within one hour, all times s_{ij} are expressed as fractions of an hour. The constraints (1d) ensure that the busy probability ζ_i is defined on $[0, 1]$ and take value 0 if no MTF is operational at i. The expected number λ_i of evacuations carried out in one hour by air ambulance i is defined by (1e). The constraints (1f) prevent from servicing any MEDEVAC from an MTF where no air ambulance is stationed, and upper-bound (M_{ij}) the number of requests at j serviced from i. The constraints (1g) ensure that all requests are serviced. Combined with (1c) and (1d), it means that all casualties are evacuated within one hour (with probability p), thereby satisficing the Golden Hour doctrine.

2.2 Boolean-Based Reformulations

The scenario-based reformulation of the chance constraint is not tenable since it includes a large number of **nonconvex constraints** equal to $|\mathbf{I}| \times |\mathbf{\Omega}|$ and of binary variables ($|\Omega|$). This motivates the development of an alternative reformulation in which the numbers of nonconvex polynomial constraints and binary variables do not increase monotonically with the number of scenarios. We extend the Boolean framework (Lejeune, 2024) to reformulate the nonlinear joint chance constraints involving endogenous uncertainty.

The Boolean approach involves three main steps: (1) construction of set of recombinations, (2) binarization of probability distribution, and (3) representation of the feasible area of the chance constraint as a partially defined Boolean function (pdBf). Using the truth table of the pdBf, one can then extract a set of mixed-integer constraints (Theorem 1) representing the feasible region.

Theorem 1 *Let $\bar{\ell}^k(i) = \max\{1 \leq l \leq n_i \mid \beta^k_{il} = 1\},\ i \in I, k \in K^-$ and let $o_{il}, i \in I, 1 \leq l \leq n_i$ be parameters measuring the distance between two consecutive cut points c_{il} and $c_{i(l+1)}$. The chance-constrained problem* **M1** *with endogenous and exogenous uncertainty can be reformulated as the MINLP problem:*

$$\mathbf{R1}: \quad \max \sum_{i \in I} y_i : \quad (x, y, z, \lambda, \zeta) \in \mathcal{MILP}$$

$$\mathcal{BLP} \begin{cases} \sum_{i\in I} \gamma_{i\bar{\ell}^k(i)+1} \geq 1 & k \in K^- & (2a) \\ \gamma_{i,l-1} \geq \gamma_{il} & i \in I,\ 2 \leq l \leq n_i & (2b) \\ \gamma_{i1} = 1 & i \in I & (2c) \\ \gamma_{il} \in \{0, 1\} & i \in I,\ 1 \leq l \leq n_i & (2d) \end{cases}$$

$$\lambda_i(1 - \zeta_i) \geq y_i \left(\sum_{l=1}^{n_i} o_{il}\gamma_{il} \right) \qquad i \in I \tag{2e}$$

It is **critical** to note that the Boolean reformulation **R1** contains |**I**| nonconvex constraints which is **several orders of magnitude less** than the (|**I**| × |**Ω**|) nonconvex constraints in the scenario-based reformulation.

Using the McCormick approach, one can reduce the reformulate the number of bilinear terms in (2e). We replace each bilinear term $y_i\gamma_{il}$ by the auxiliary variable w_{il} defined by the linear inequalities in $\mathcal{MC}^1_{il} := \{w_{il} \geq 0,\ w_{il} \geq u^y_i\gamma_{il} + y_i - u^y_i,\ w_{il} \leq u^y_i\gamma_{il},\ w_{il} \leq y_i\}$; (2e) can be rewritten as: $\lambda_i(1-\zeta_i) \geq \sum_{l=1}^{n_i} o_{il}w_{il},\ i \in I$.

Theorem 2 presents another nonconvex quadratic integer optimization problem **R2** which reduces the number of nonseparable bilinear terms in $\lambda_i(1 - \zeta_i)$, and linearizes some of the squared bilinear terms in $\lambda_i(1 - \zeta_i)$. We rewrite (2e) in its extensive form using (1c) and (1e). Let $H_i \subseteq J$ designate the subset of CCP locations that can be serviced from MTF i within acceptable time. vspace-0.2in

Theorem 2 *Define the parameters* $s_{ij+j'} = s_{ij}+s_{ij'}, j \in H_i, i \in I$. *Let* $G_{ij} := \{j' \in H_i\backslash\{j\} : s_{ij+j'} = s_{ij}+s_{ij'} \leq 1\} \subseteq H_i,\ j \in H_i, i \in I,\ F^m_{ij} := \{j' \in G_{ij} : j' > j\} \subseteq G_{ij},\ j \in H_i, i \in I$ *and* $F^1_i = \{j \in H_i : u^x_{ij} = 1\} \subseteq H_i, i \in I$. *The MINLP problem* **R2** *is equivalent to* **R1**

$$\mathbf{R2} : \max \sum_{i\in I} y_i :\ (x, y, z, \lambda, \zeta, \gamma, w) \in \mathcal{MILP}\ \cap\ \mathcal{BLP}\ \cap\ \mathcal{MC}^1_{il}$$

$$\lambda_i - \sum_{j\in H_i} \sum_{j'\in F^m_{ij}} s_{ij+j'}x_{ij}x_{ij'} - \sum_{j\in F^1_i} s_{ij}x_{ij} - \sum_{j\in H_i\backslash F^1_i} s_{ij}(x_{ij})^2 \geq \sum_{l=1}^{n_i} o_{il}w_{il}\ ,\ i \in I \tag{3}$$

and has the minimum number of bilinear nonseparable terms and of squared bilinear terms.

2.3 Algorithmic Method

We design a nonlinear BB algorithm which differs from traditional nonlinear BBs in two main ways. First, we design a **conic relaxation** for the nonconvex MINLP problem and solve at each node of the BB tree a conic mixed-integer problem instead of a continuous polyhedral relaxation. Second, the choice of branching scheme, called **smallest domain branching rule**, is unconventional.

Conic integer bounding and solution of conic mixed-integer subproblems: For each constraint (3) in **R2**, we derive a multiterm conic relaxation on the **minimal collection of nonseparable, nonconvex terms** that can take a positive value and leave unchanged the linear and quadratic convex terms in (3). The mixed-integer bounding problem **B2** is foundational for the BB algorithm as **B2** can be solved very quickly.

Theorem 3 *Let* $S^{min}_i = \min_{j\in H_I, j'\in F^m_{ij}} s_{ij+j'}, i \in I$. *The conic MINLP problem* **B2**

$$\mathbf{B2} :\ \max \sum_{i\in I} y_i : (x, y, z, \lambda, \zeta, \gamma, w) \in \mathcal{MILP}\ \cap\ \mathcal{BLP}\ \cap\ \mathcal{MC}^1_{il}$$

$$\lambda_i - v_i \geq \sum_{l=1}^{n_i} o_{il}w_{il} \quad , \quad v_i \geq \sum_{j\in H_i\backslash F^1_i} s_{ij}(x_{ij})^2 + t_i + \sum_{j\in F^1_i} s_{ij}x_{ij} \qquad i \in I \tag{4a}$$

$$t_i \geq (d_i - 1)S^{min}_i \quad , \quad \sum_{j\in F^1_i} x_{ij} + \sum_{j\in H_i : u^x_{ij}=2} x_{ij_1} = d_i \qquad i \in I \tag{4b}$$

is a relaxation of **R2**. *Its optimal value is an upper bound on the optimal value of* **M1**.

Consider the auxiliary variables v_i defined as $v_i := \zeta_i\lambda_i, i \in I$ in **R2**. We can replace these nonlinear equalities by the set $\mathcal{MC}^2_i$ of mixed-integer linear McCormick inequalities which defines the convex and concave envelope of each individual bilinear term $v_i := \zeta_i\lambda_i$. A key departure from what MINLP solvers do is that, at each node of the tree, we solve the **mixed-integer conic relaxation problem**:

$$\mathbf{B2\text{-}MC} :\ \max \sum_{i\in I} y_i :\ (x, y, z, \lambda, \zeta, \gamma, w, v) \in \mathcal{MILP}\ \cap\ \mathcal{BLP}\ \cap\ \mathcal{MC}^1_{il}\ \cap\ \mathcal{MC}^2_i\ \cap\ (4a) - (4b) \tag{5}$$

Any feasible solution for **B2-MC** satisfying all constraints $v_i = \zeta_i\lambda_i, i \in I$ yields a feasible solution to **M1**. Additionally, the solutions of **B2-MC** and **M1** have the same objective value. Hence, a solution satisfying all constraints $v_i = \zeta_i\lambda_i$ with the best possible objective value for **B2-MC** yields an optimal solution to **M1**.

Branching rule: The second difference is the new branching method, called *smallest domain branching rule* (SDBR). Let **B2-MC**(k) be the problem **B2-MC** using the lower and upper bounds valid at node k of the branch-and-bound tree and let $(\zeta^k, \lambda^k, v^k)$ be one of its optimal solution. If $(\zeta^k, \lambda^k, v^k)$ satisfies all constraints $v_i := \zeta_i \lambda_i$, it is saved as an incumbent and node k is pruned. Otherwise, we select an index $i(B) \in I$ corresponding to the most violated constraint $v_i = \zeta_i \lambda_i$: $i(B) = \text{argmax}_{i \in I} |v_i^k - \zeta_i^k \lambda_i^k|$. We branch on $\lambda_{i(B)}$ or $\zeta_{i(B)}$ by splitting the range and updating the McCormick inequalities in the two corresponding subproblems.

3 Results

We use real-life data about the Operation Enduring Freedom in Afghanistan where fourteen air ambulances were used for MEDEVACs. The theatre of operations is the NATO Regional Command-South and covers several Afghan provinces. We use Baron to solve the nonconvex MINLP problems **R1** and **R2** and Cplex to solve **B2**. We have created 192 realistic problem instances which differ in terms of number of potential sites for MTFs, number of CCPs, probability level, and congestion C level.

3.1 Computational Tractability of Reformulations

We assess the tractability of the reformulations **R1** and **R2**. Note that **none** of the problem instances modelled with the scenario-based approach could be solved in 10 hours. Table 1 reports the average (μ^t), smallest (l^t), and largest (u^t) solution time across the 192 instances (last row). An entry with “3600" signifies that optimality could not be proven in one hour. The superiority of the nonconvex MINLP reformulation **R2** is manifest. It can prove optimality for each instance in less than one hour whereas **R1** fails to do so for 61 instances: **R2** is much faster than **R1** and takes less than 15 minutes for all but 3 instances. This highlights the value of minimizing the number of bilinear terms, which differentiates **R1** and **R2**.

3.1.1 Tightness and Solution Times for Conic Bounding Problem

The upper bounds obtained from the relaxation problem **B2** are remarkabley tight as the average optimality gap across the 192 instances is 0.14% for **B2**. Furthermore, problem **B2** can be solved very quickly. The average and largest solution times for **B2** are respectively equal to 0.19 and 9.77 seconds.

3.1.2 Efficiency of Nonlinear BB Algorithm

We now integrate the relaxation problem **B2** within the nonlinear BB algorithmic framework. The enumeration tree itself is coded using the Bcp framework. The results show unequivocally that the nonlinear BB algorithm B&B2 is extremely fast: (i) the average solution time across the 192 instances is 1.28 seconds and (ii) B&B2 can solve and prove optimality for optimality in no more than 27.2 seconds.

Table 1: Tractability of Reformulations (sec)

Instance Type		μ^t (sec.)		$[l^t, u^t]$ (sec.)	
		R1	**R2**	**R1**	**R2**
p	0.95	496.83	61.10	[0.05 , 3600]	[0.30 , 1170.37]
	0.925	1520.21	69.00	[0.96 , 3600]	[0.59 , 1357.26]
	0.90	1925.58	111.31	[0.36 , 3600]	[0.38 , 822.49]
C	I	1268.34	54.05	[0.08 , 3600]	[0.38 , 746.81]
	II	1335.10	60.28	[0.07 , 3600]	[0.64 , 673.54]
	III	1406.59	97.61	[0.05 , 3600]	[0.30 , 1357.26]
	IV	1276.22	114.62	[0.09 , 3600]	[0.54 , 1170.37]
All Instances		1321.57	81.07	[0.05 , 3600]	[0.30 , 1357.54]

Table 2: Efficiency of Nonlinear BB

Instance Type		μ^t (sec.)	$[l^t, u^t]$ (sec.)	$\tilde{m}$
p	0.95	0.88	[0.0 , 11.1]	0
	0.925	0.82	[0.1 , 7.3]	0
	0.90	2.16	[0.0 , 27.2]	0
C	I	0.75	[0.0 , 11.1]	0
	II	1.32	[0.0 , 27.2]	0
	III	1.53	[0.0 , 25.8]	0
	IV	1.51	[0.0 , 26.2]	0
All Instances		1.28	[0.00 , 27.2]	0

4 Discussion

We propose a new chance-constrained MEDEVAC problem with exogenous and endogenous uncertainties. The model incentivizes the fastest possible air ambulances-based MEDEVAC and provides a reliability guarantee to do so. We develop a general Boolean reformulation approach and a nonlinear BB algorithm which features: 1) a new multiterm convexification approach for polynomial constraints, 2) the solution of a conic relaxation problem at each node, and 3) the new SDBR branching rule. The tests attest the computational efficiency and applicability of the approach, even for highly congested MEDEVAC problems.

References

Erkut, E., Ingolfsson, A., & Erdogan, G. 2008. Ambulance Location for Maximum Survival. *Naval Research Logistics*, **55**, 42–58.

Lejeune, Miguel A. 2024. Boolean Reformulation Method for Linear and Nonlinear Joint Chance Constraints. *In:* Pardalos, Panos M, & Prokopyev, Olege A (eds), *Encyclopedia of Optimization*. Springer.

Shackelford, S.A., Del Junco, D., Mazuchowski, E.L., Kotwal, R.S., Remley, M.A., Keenan, S., & Gurney, J.M. 2024. The Golden Hour of Casualty Care: Rapid Handoff to Surgical Team is Associated With Improved Survival in War-Injured US Service Members. *Annals of Surgery*, **1**, 1–10.