International Journal of Applied Operational Research
Vol. 14, No. 3, pp. 1-16, Summer 2026

Journal homepage: ijorlu.liau.ac.ir

Cost efficiency and congestion analysis in pharmaceutical firms: A two-stage DEA framework

M. Outougari, F. Hosseinzadeh Lotfi*, M. Rostamy-Malkhalifeh, T. Allahviranloo

Received: 21 April 2026 ; **Accepted:** 6 June 2026

Abstract This study proposes a two-stage integrated data envelopment analysis (DEA) framework to evaluate cost efficiency and identify cost congestion in decision-making units (DMUs). The framework reflects real production conditions by considering discretionary and non-discretionary inputs, as well as desirable and undesirable outputs. In the first stage, a cost-oriented DEA model measures the cost efficiency of DMUs while accounting for non-discretionary inputs. Cost-efficient units are identified and used as benchmark reference units. In the second stage, a diagnostic procedure is introduced to detect and quantify cost congestion in inefficient DMUs relative to the efficient benchmarks. Cost congestion refers to situations where excessive use of discretionary inputs leads to reduced output performance. The framework is applied to pharmaceutical companies as an empirical case. The results show that cost inefficiency is not always due to poor management, and some inefficient DMUs experience significant cost congestion. In such cases, increasing inputs worsens performance instead of improving it. The magnitude of congestion is explicitly quantified, enabling a clear distinction between managerial inefficiency and structural congestion. The proposed approach provides a benchmark-based method for congestion identification and integrates cost efficiency and congestion analysis into a unified DEA framework. This contributes to the DEA literature under heterogeneous production conditions.

Keyword: Cost Efficiency, Data Envelopment Analysis, Cost Congestion, Non-Discretionary Inputs, Undesirable Outputs.

1. Introduction

Data envelopment analysis (DEA) is a well-established non-parametric methodology for evaluating the relative performance of a set of homogeneous decision-making units (DMUs) that employ multiple inputs to produce multiple outputs. Since the seminal work of Charnes et

* **Corresponding Author.** (✉)
E-mail: Farhad@hosseinzadeh.ir (F. Hosseinzadeh Lotfi)

M. Outougari
Department of Mathematics, SR.C., Islamic Azad University, Tehran, Iran

F. Hosseinzadeh Lotfi
Department of Mathematics, SR.C., Islamic Azad University, Tehran, Iran

M. Rostamy-Malkhalifeh
Department of Mathematics, SR.C., Islamic Azad University, Tehran, Iran

T. Allahviranloo
Department of Mathematics, SR.C., Islamic Azad University, Tehran, Iran

[DOI: 10.71885/ijorlu-2026-3-733]

al. (1978), which introduced the CCR model under constant returns to scale (CRS), DEA has undergone substantial methodological development [1]. The subsequent extension by Banker et al. (1984) to variable returns to scale (VRS) significantly broadened the applicability of DEA in empirical studies. Over the past decades, DEA has evolved into a mature benchmarking framework and has been widely applied across various sectors, including banking, healthcare, manufacturing, energy, and education [2,3]. Recent reviews highlight that DEA continues to expand with integrations of machine learning, robust optimization, and sustainability metrics, reflecting its adaptability to complex real-world problems [4].
An important extension of traditional technical efficiency analysis within the DEA framework is cost efficiency (CE). Cost efficiency evaluates a DMU's ability to produce a given bundle of outputs at the minimum possible cost, given prevailing input prices. Unlike technical efficiency, which focuses solely on physical input–output relationships, cost efficiency explicitly incorporates economic behavior and allocative considerations [5,6]. DEA-based cost efficiency models have therefore become indispensable tools in complex production environments where specifying explicit cost functions is impractical [7]. Recent studies have further refined cost efficiency analysis by addressing imperfect price information, regulatory constraints, statistical inference issues, and integration with environmental and sustainability goals [8].

Beyond technical and allocative inefficiency, an additional and more subtle source of performance deterioration is congestion. First formally introduced by Färe and Svensson (1980), congestion refers to a situation in which an increase in one or more inputs leads to a reduction in one or more outputs, holding other factors constant. Congestion represents a severe form of diseconomy that is fundamentally different from scale inefficiency or pure managerial inefficiency. Methodological advances in recent years have revisited congestion measurement within the DEA framework. These include approaches under multiple optimal solutions, directional distance functions, network structures, and robust models that better handle uncertainty and heterogeneity in production systems [9,10].

In practical applications, accurately identifying congestion is complicated by two pervasive characteristics of real-world production systems: the presence of non-discretionary inputs (such as regulatory conditions, inherited infrastructure, or external economic environments like exchange rates that are beyond managerial control) and the joint production of undesirable outputs (such as emissions, waste, or financial liabilities). Banker and Morey (1986) pioneered the treatment of non-discretionary inputs, while Färe et al. (1989) and Seiford and Zhu (2002) laid foundational work for undesirable outputs under weak disposability [11,12]. Recent contributions have advanced these areas by developing models that jointly handle non-discretionary factors and undesirable outputs in efficiency and congestion analysis, often incorporating stochastic elements, super-efficiency, or sustainability-oriented frameworks [13-16].

Although numerous DEA extensions have been proposed to address non-discretionary inputs and undesirable outputs separately, their joint incorporation into cost congestion analysis remains limited. Cost congestion arises when additional discretionary expenditures increase total cost without improving (and sometimes deteriorating) outputs, directly threatening financial sustainability. Misinterpreting cost congestion as standard inefficiency may lead to counterproductive managerial decisions, such as allocating additional budgets to units that require structural redesign rather than more resources [17,18].

To address this gap, this study proposes a two-stage integrated DEA framework for cost efficiency assessment and cost congestion diagnosis. In the first stage, a cost-oriented DEA model is employed to evaluate the cost efficiency of DMUs while explicitly accounting for

[DOI: 10.71885/ijorlu-2026-3-733]

non-discretionary inputs. Cost-efficient DMUs are identified and used as reference units. In the second stage, a diagnostic algorithm is introduced to detect and quantify cost congestion in inefficient DMUs by benchmarking them against the set of cost-efficient units. By simultaneously accommodating discretionary and non-discretionary inputs as well as desirable and undesirable outputs, the proposed framework advances DEA-based performance evaluation from descriptive efficiency measurement toward prescriptive managerial diagnosis.

This study makes several important contributions. First, it proposes a novel two-stage DEA framework that integrates cost efficiency and cost congestion under heterogeneous production conditions. Second, it explicitly distinguishes structural cost congestion from general managerial inefficiency. Third, the empirical application to Iranian pharmaceutical companies demonstrates the practical value of the model in resource-intensive industries facing high costs, currency fluctuations, and regulatory pressures. The results provide actionable insights for managers and policymakers aiming to enhance both economic and environmental performance.

2. Background

2.1. Data envelopment analysis and the production possibility set (PPS)

DEA is a sophisticated branch of Operations Research that evaluates the efficiency of DMUs, encompassing a wide range of entities such as banks, universities, and corporations. The performance of these units is directly influenced by the resources they consume, where, in most cases, an increase in input levels results in higher output. However, this relationship is not universally applicable across all production environments, as certain technological constraints may disrupt the expected input-output proportionality. To assess efficiency, DEA constructs a *PPS* a set that delineates all feasible production activities based on empirical observations. The efficient frontier, a crucial subset of the PPS boundary, serves as an approximation of the underlying production function. Each DEA model is characterized by a distinct PPS, which is uniquely determined by its foundational assumptions and underlying axiomatic principles.

Suppose we have n DMUs with m inputs and s outputs, and that the vectors $x_j = \left(x_{1j},...,x_{mj}\right)^T$ and $y_j = \left(y_{1j},...,y_{sj}\right)^T$ denote the input and output values of DMU_j, $j=1,...,n$, respectively. Additionally, at least one of the inputs and outputs is non-zero $(x_j \underset{\neq}{\geq} 0\ ,\ y_j \underset{\neq}{\geq} 0)$.

The production technology may adhere to the following principles:

(1) Observation Envelopment: $\left(x_j, y_j\right) \in T$, $j=1,...,n$.

(2) Convexity: if $\left(x_1,y_1\right),\left(x_2,y_2\right) \in T$ and $\lambda \in [0,1]$ then, $\lambda\left(x_1,y_1\right)+(1-\lambda)\left(x_2,y_2\right) \in T$

(3-1) Output possibility: if $\left(x_1,y_1\right) \in T$, and $y_2 \le y_1$ then $\left(x_1,y_2\right) \in T$.

(3-2) Input possibility: if $\left(x_1,y_1\right) \in T$, and $x_2 \ge x_1$ then $\left(x_2,y_1\right) \in T$.

(4) Ray Unboundedness: if $\left(x, y\right) \in T$ and $\alpha \ge 0$, then, $\left(\alpha x, \alpha y\right) \in T$

(5) Minimum Interpolation: By accepting this principle, we acknowledge that T is the smallest set that satisfies the first through fourth principles.

The production possibility set that satisfies principles 1, 2, 3-1, and 5, denoted by T_v or T_{BCC}, is represented as follows:

[DOI: 10.71885/ijorlu-2026-3-733]

$$T_v = T_{BCC} = \left\{ (x, y) \middle| x \geq \sum_{j=1}^{n} \lambda_j x_j,\ y \leq \sum_{j=1}^{n} \lambda_j y_j, \sum_{j=1}^{n} \lambda_j = 1,\ \lambda_j \geq 0\ j = 1, \ldots, n \right\}$$

2.2. Congestion in DEA and the Revised PPS

In some scenarios, increasing input resources can paradoxically lead to reduced output. For instance, hiring additional employees in an office environment may lead to congestion, where excessive staff congestion results in operational inefficiencies. Employees may unintentionally obstruct one another's tasks, decreasing both the quality and quantity of services rendered. This phenomenon, known as congestion**,** arises when an excessive concentration of resources within a constrained system leads to diminished productivity.

Definition 1. Congestion is said to occur when the output that is maximally possible can be increased by reducing one or more inputs without improving any other inputs or outputs. Conversely, congestion is said to occur when some of the outputs that are maximally possible are reduced by increasing one or more inputs without improving any other inputs or outputs.

To accurately model congestion, the traditional PPS must be revised. Unlike standard DEA frameworks, where input feasibility is assumed, the congestion-adjusted PPS does not adhere to the input feasibility axiom. This distinction allows for a more precise representation of scenarios in which resource overutilization adversely impacts efficiency.

The production possibility set that satisfies principles 1, 2, 3, and 5, denoted by T_N, is represented as follows:

$$T_N = \left\{ (x, y) \middle| x = \sum_{j=1}^{n} \lambda_j x_j,\ y \leq \sum_{j=1}^{n} \lambda_j y_j, \sum_{j=1}^{n} \lambda_j = 1,\ \lambda_j \geq 0\ j = 1, \ldots, n \right\}$$

2.3. The evolution of congestion measurement in DEA

The formal identification and quantification of congestion within DEA began with the FGL method, introduced by Färe (1980) [19]. This approach defined congestion through an isoquant under the weak disposability assumption of inputs, allowing for proportional input reductions or expansions. However, Cooper (2001) later identified shortcomings in the FGL method when applied under the strong disposability assumption, where input ratios could vary freely. In response, Cooper, W.W., Thompson, R.G., Thrall, R.M. (1996) introduced the *CTT method*, which operates under strong disposability, offering a refined perspective on congestion assessment [20].

Subsequent advancements came from Wei (2004) and Tone (2004), who constructed a Production Possibility Set (PPS) explicitly designed to capture congestion phenomena. This revised PPS redefined congestion as a frontier characteristic, removing the disposability assumption of inputs to enable more precise identification of overinvestment. The *WY-TS method*, derived from this PPS, evaluates whether a Decision-Making Unit (DMU) can reduce inputs while increasing outputs, thereby detecting congestion. Within this framework, congestion manifests in two distinct forms [21]:

- Strong congestion**,** where reducing all inputs leads to an increase in all outputs.
- Weak congestion, which aligns with the conditions set forth in Definition 1.

[DOI: 10.71885/ijorlu-2026-3-733]

Subsequent research has further refined congestion measurement methodologies. Cooper, (2002) proposed a *one-step congestion identification method*, which was later extended into stochastic DEA and fuzzy DEA frameworks. Noura, (2010) introduced a more streamlined approach, known *as* Noura's method, which builds upon the CTT framework while simplifying congestion detection [22].

2.4. Cost efficiency

Assume that a cost can be assigned to each input. Let c_i be the cost of one unit of input x_{ip}. In this case, the input vector can be transformed into a single aggregated input, denoted as $CX_p \in \mathbb{R}$.

Thus, under this assumption, each decision-making unit (DMU) has a single input representing the total cost and an output vector, which can be expressed as follows:

$$(CX_j, Y_j) \in \mathbb{R}^{1+s}_{\geq 0}$$

Based on the following Production Possibility Set**,**

$$T_c = \left\{ \begin{pmatrix} X \\ Y \end{pmatrix} \middle| X \geq \sum_{j=1}^{n} \lambda_j X_j \;\&\; Y \leq \sum_{j=1}^{n} \lambda_j Y_j \;\&\; \lambda \geq 0 \right\}$$

the cost efficiency model is formulated as follow**s:**

$$\begin{aligned} &Min \quad CX \\ &st: \begin{pmatrix} X \\ Y_p \end{pmatrix} \in T_c \end{aligned}$$

Cost efficiency (CE_P)is defined as the ratio of the minimum achievable cost to the existing cost, with its value always falling within the interval [0,1].

Definition2: DMU_p is called cost efficient if and only if $CE_P = 1$.

2.5. The Noura's method

The Noura's method involves three distinct steps:

Step 1. Identify BCC-efficiency using Model (1).

$$\max \;\; \varphi_o + \varepsilon\left(es_o^- + es_o^+\right)$$

$$s.t.\begin{cases} \sum_{j=1}^{n} \lambda_j x_j = x_o - s_o^-, \\ \sum_{j=1}^{n} \lambda_j y_j = \varphi_o y_o + s_o^+, \\ \sum_{j=1}^{n} \lambda_j = 1, \\ (\lambda, s_o^-, s_o^+) \geq 0 \end{cases} \tag{1}$$

The variable φ_o represents the maximum feasible expansion of desirable outputs. A non-Archimedean infinitesimal (ε) is employed theoretically to avoid re-specifying the model. In

[DOI: 10.71885/ijorlu-2026-3-733]

practice, this necessitates solving two separate models, each with its own objective function but identical constraints, to determine the optimal solution. e indicates the vector $[1,1,...,1]$, and 0 represents the vector $[0,0,...,0]^T$. The vector λ_j denotes the intensity levels at which the DMU can produce. The constraint $\sum_{j=1}^{n} \lambda_j = 1$ ensures variable returns to scale. Slacks for inputs and outputs are represented by s_o^-, s_o^+, respectively.

Solve Model (1) for each DMU_0. If $\varphi_o^* = 1$ and $s_o^{+*}, s_o^{-*} = 0$ (input and output slacks are zero, indicating no excess inputs or outputs) then DMU_0 is efficient. BCC-efficient $DMUs$ form set $E = \left\{ j \,\middle|\, \varphi_j^* + \varepsilon\left(es_j^{-*} + es_j^{+*}\right) = 1 \right\}$, where $*$ denotes optimality in Model (1).

Step 2. Among the $DMUs$ in the set E, there exists at least one, say DMU_l, that has the highest consumption in its first input component compared with the first input component of the remaining $DMUs$ of set E. Let:

$$x_1^* = x_{1l} = Max\left\{ x_{1j} : j \in E \right\} \qquad \forall j \ (j \in E \Rightarrow x_1^* \geq x_{1j})$$

Next, a DMU, denoted as DMU_k, is identified within the set E that has the highest input value in its second component when compared to other $DMUs$ in the same set.

$$x_2^* = x_{2k} = Max\left\{ x_{2j} : j \in E \right\} \qquad \forall j \ (j \in E \Rightarrow x_2^* \geq x_{2j})$$

This process is repeated similarly for all components of the inputs, ultimately resulting in:

$$x^* = \left(x_1^*, x_2^*, ..., x_m^* \right)$$

Step 3. Identify DMU_o as congested if any of the following conditions hold:

(1) $\varphi_o^* > 1$, where * indicates the optimal solution of Model (1) and there is at least one $x_{io} > x_i^*$ for $i = 1,...,m$;

(2) $es_o^{+*} > 0$, where * indicates the optimal solution of Model (1) and there is at least one $x_{io} > x_i^*$ for $i = 1,...,m$.

The amount of excessive part of input x_{io} can be calculated by $s_{io}^c = x_{io} - x_i^*$.

This method significantly reduces computational complexity by streamlining the identification process of congested inputs. Despite the substantial body of research on DEA-based cost efficiency and congestion analysis, Table 1 demonstrates that existing studies address these issues in a fragmented manner and fail to provide an integrated framework that simultaneously considers cost-based congestion, non-discretionary inputs, and undesirable outputs.

Table 1. summary of related Studies and identified research gaps in DEA-based cost congestion analysis

Study	Cost Efficiency	Congestion Analysis	Cost-Based Congestion	Non-Discretionary Inputs	Undesirable Outputs	Integrated Two-Stage Framework
Färe & Grosskopf (1985) [6]	❌	✅	❌	❌	❌	❌
Cooper et al. (2001) [20]	❌	✅	❌	❌	❌	❌

[DOI: 10.71885/ijorlu-2026-3-733]

Wei & Yan (2004) [21]	✗	✅	✗	✗	✗	✗
Tone & Sahoo (2004) [10]	✗	✅	✗	✗	✗	✗
Noura et al. (2010) [22]	✗	✅	✗	✗	✗	✗
Liu et al. (2015) [15]	✗	✗	✗	✗	✅	✗
Emrouznejad et al. (2019) [3]	✅	✗	✗	✗	✅	✗
Tavassoli et al. (2020) [23]	✗	✗	✗	✗	✅	✗
Yousefi et al. (2023) [24]	✗	✅	✗	✗	✗	✗
This study	☑	☑	☑	☑	☑	☑

Table 1 summarizes the main methodological features of representative studies in the DEA literature on cost efficiency and congestion analysis. As shown, existing approaches predominantly address cost efficiency or congestion effects in isolation. Although some studies incorporate non-discretionary inputs or undesirable outputs, these features are rarely considered jointly, particularly within a cost-based congestion framework. In particular, none of the reviewed studies provides a unified framework that simultaneously integrates cost efficiency evaluation, cost congestion diagnosis, non-discretionary inputs, and undesirable outputs within a benchmark-driven, two-stage structure. The present study fills this methodological gap by proposing a novel integrated two-stage DEA framework that explicitly distinguishes structural cost congestion from general cost inefficiency. This integration enables more accurate performance assessment and provides enhanced diagnostic capability for supporting efficient and sustainable resource allocation decisions.

3. Proposed Method

This study aims to calculate the cost efficiency of the evaluated DMUs in the presence of both discretionary and non-discretionary inputs, as well as desirable and undesirable outputs, using a cost efficiency model. Subsequently, by applying the algorithm proposed in the following sections, it identifies and resolves cost congestion for those units found to be cost-inefficient.
We assume there are n DMUs, where each DMU consumes a vector of discretionary and non-discretionary inputs, denoted by $(X_j \in \mathbb{R}^I, X_j^N \in \mathbb{R}^M)$, to produce a vector of desirable and undesirable outputs, denoted by $(Y_j^g \in \mathbb{R}^S, Y_j^b \in \mathbb{R}^T)$. By accepting the standard axioms of production theory namely, inclusion of observations**,** convexity**,** input and output feasibility**,** *and* minimum interpolation we define the production possibility set PPS as follows:

$$T_V = \left\{ \left(X, X^N, Y^g, Y^b\right) \middle| \begin{array}{ll} \sum_{j=1}^{n} \lambda_j x_{ij} \le x_i & i=1,\dots,I \\ \sum_{j=1}^{n} \lambda_j x_{mj}^N) \le x_{mp}^N & m=1,\dots,M \\ \sum_{j=1}^{n} \lambda_j \left(y_{rj}^g \theta_j\right) \ge y_{rp}^g & r=1,\dots,s \\ \sum_{j=1}^{n} \lambda_j \left(y_{tj}^b \theta_j\right) = y_{tj}^b & t=1,\dots,T \\ \sum_{j=1}^{n} \lambda_j = 1 & \\ \theta_j \in [0,1] \ , \ j=1,\dots,n \ , \ \lambda_j \ge 0 & \end{array} \right\} \quad (2)$$

Based on the convexity assumption, the variable $\lambda_j, (j=1,\dots,n)$ are used in the structure of the technology T_V as the intensity weights to from the linear combination of n observed $DMUs$. The assumption of VRS is maintained by restriction that the sum of these intensity variables is 1, i.e. $\sum_{j=1}^{n} \lambda_j = 1,$ and θ_j is the endogenous abatement factor (Epure M, Lafuente E 2015), which allows the simultaneous adjustments of both desirable and undesirable outputs.
To transform 2 into a linear form, we let $\lambda_j \theta_j = \mu_j$ and $v_j = \left(1-\theta_j\right)\lambda_j = \lambda_j - \mu_j$, then $\mu_j \ge 0, v_j \ge 0$, $\lambda_j = \mu_j + v_j \ge 0$ and $\sum_{j=1}^{n} \lambda_j = \sum_{j=1}^{n} (\mu_j + v_j) \ge 0$ hold.

$$T_v = \left\{ \left(X, X^N, Y^g, Y^b\right) \middle| \begin{array}{ll} \sum_{j=1}^{n} \left(\mu_j + v_j\right) x_{ij} \le x_i & i=1,\dots,I \\ \sum_{j=1}^{n} \left(\mu_j + v_j\right) x_{mj}^N \le x_{mp}^N & m=1,\dots,M \\ \sum_{j=1}^{n} \mu_j y_{rj}^g \ge y_{rp}^g & r=1,\dots,s \\ \sum_{j=1}^{n} \mu_j y_{tj}^b = y_{tp}^b & t=1,\dots,T \\ \sum_{j=1}^{n} \left(\mu_j + v_j\right) = 1 \quad , \quad \mu, v \ge 0 & \end{array} \right\} \quad (3)$$

Based on the production possibility set defined above, we define the cost efficiency model as follows:

[DOI: 10.71885/ijorlu-2026-3-733]

$$
\begin{aligned}
P = Min \ & \frac{\sum_{i=1}^{I} c_i x_i}{\sum_{i=1}^{I} c_i x_{ip}} \\
s.t.: \ & \sum_{j=1}^{n} \left(\mu_j + \nu_j\right) x_{ij} \le x_i && i = 1,\ldots,I \\
& \sum_{j=1}^{n} \left(\mu_j + \nu_j\right) x_{mj}^{N} \le x_{mp}^{N} && m = 1,\ldots,M \\
& \sum_{j=1}^{n} \mu_j y_{rj}^{g} \ge y_{rp}^{g} && r = 1,\ldots,s \\
& \sum_{j=1}^{n} \mu_j y_{tj}^{b} = y_{tp}^{b} && t = 1,\ldots,T \\
& \sum_{j=1}^{n} \left(\mu_j + \nu_j\right) = 1 \\
& \mu_j, \nu_j \ge 0 \quad X \ge 0 && j = 1,\ldots,n
\end{aligned} \tag{4}
$$

$$
T_{New} = \left\{ \left(X, X^N, Y^g, Y^b\right) \middle| \begin{array}{ll}
\sum_{j=1}^{n} \left(\mu_j + \nu_j\right) x_{ij} = x_i & i = 1,\ldots,I \\
\sum_{j=1}^{n} \left(\mu_j + \nu_j\right) x_{mj}^{N} \le x_{mp}^{N} & m = 1,\ldots,M \\
\sum_{j=1}^{n} \mu_j y_{rj}^{g} \ge y_{rp}^{g} & r = 1,\ldots,s \\
\sum_{j=1}^{n} \mu_j y_{tj}^{b} = y_{tp}^{b} & t = 1,\ldots,T \\
\sum_{j=1}^{n} \left(\mu_j + \nu_j\right) = 1 \ , \ \mu, \nu \ge 0 &
\end{array} \right\} \tag{5}
$$

Definition (Cost Congestion): DMU_p exhibits cost congestion if and only if the following conditions are simultaneously satisfied:

1) With an increase in CX_p by ΔCX_p, the value of Y_p^g decreases for at least one index in T_{New} *or* the value of y_p^b increases for at least one index .
2) With a decrease in CX_p by ΔCX_p, the value of Y_p^g increases for at least one index in T_{New}

Algorithm:

- First, all units are evaluated using Model 3 to calculate their cost efficiency. Subsequently, those units identified as efficient by the model are placed in a set denoted as *Set E*. $E = \left\{ j \middle| p_j^* = 1 \right\}$

[DOI: 10.71885/ijorlu-2026-3-733]

- Next, from among all units in Set E, we select the unit(s) with the highest value of discretionary cost input(s), defined as follows: $\overline{CX} = Max\left\{CX_j \,\middle|\, j \in E\right\}$
- DMU_p under evaluation is considered to have cost congestion in its discretionary input if and only if the following two conditions hold simultaneously:
  1) $\rho_p^* < 1$
  2) $CX_p > \overline{CX} \qquad p \notin E$

Theorem 1: If $CX_P < \overline{CX}$, then DMU_p has no cost congestion.

Proof: If DMU_p is cost-effective, according to item 2 of the definition of cost congestion, since the cost of discretionary inputs cannot be reduced, it therefore lacks cost congestion. If DMU_p is inefficient in terms of cost, then assume $\overline{CX_p} = CX_t$. That is, DMU_t has the highest cost but is still efficient. Therefore, by increasing CX_p by $CX_t - CX_p > 0$, there exists a point in T_v where its outputs are no worse than those of DMU_p (since DMU_t is cost-efficient). Hence, according to item 1 of the definition, DMU_p lacks cost congestion.

Theorem 2: DMU_p is considered to have cost congestion if the following conditions hold. Its cost congestion measure (or score) is equal to $CX_p - \overline{CX}$.

1) $CX_p = Max\left\{CX_j \,\middle|\, j = 1,...,n\right\}$

2) $CX_p > \overline{CX}$

3) If DMU_p is not efficient in output-oriented in Tv.

Proof :

- State *A:* If $CX_p = Max\left\{CX_j \,\middle|\, \forall j\right\}$, then if $CX_p \to CX_p + \Delta CX_p$, a new point with coordinates $\left(CX_p + \Delta CX_p, X_p^N, Y_p^g, Y_p^b\right)$ will not exist in T_{new}. Therefore, by increasing CX_p by ΔCX_p, the desired output will not increase in T_{New}.

Now, assume we decrease CX_p by a small amount. We must prove that Y_p^g can increase.

- *State B:* If $CX_p = Max\left\{CX_j \,\middle|\, \forall j\right\}$, then we reduce the input by ΔCX_p, that is, $CX_p \to CX_p - \Delta CX_p$ is a new point with coordinates $\left(CX_p - \Delta CX_p, X_p^N, Y_p^g, Y_p^b\right) \in T_{new}$. According to Condition 3, since DMU_p is inefficient in T_{New}, there exists a DMU that dominates it.

Given the conditions

$$\overline{CX} = Max\left\{CX_j \,\middle|\, j \in E\right\}$$
$$CX_p = Max\left\{CX_j \,\middle|\, \forall j\right\}$$
$$CX_p > \overline{CX}$$

Moreover, since DMU_p belongs to T_{New}, a reduction in the input does not lead to a decrease in the desirable output; therefore, there exists an increase in the output in at least one index.

4. Empirical example and data description

This study analyzes a sample of 21 pharmaceutical firms to evaluate cost efficiency and identify the presence of cost congestion. The data were obtained from audited financial statements of pharmaceutical companies listed on the stock exchange and publicly available financial disclosure systems. The fiscal year 2018 was chosen to ensure cross-sectional consistency and comparability across firms. Capital and total assets are treated as discretionary inputs. These variables are measured in monetary terms and represent the financial value of resources employed in production and operational activities. Since these inputs are expressed in monetary units, they inherently reflect resource costs, and separate input price information is not required for the cost efficiency analysis. Foreign exchange expenditure is considered a non-discretionary input, as it is largely influenced by exchange rate fluctuations beyond managerial control. Sales and market value are considered desirable outputs. Sales represent revenue generated from core business activities, while market value reflects investors' expectations regarding firm performance and future growth. Total liabilities are treated as an undesirable output, as higher liabilities increase financial risk and reduce financial flexibility. All monetary values are reported in local currency units. A summary of the input and output variables is presented in Table 2.

Table 2. Description of inputs and outputs used in the cost congestion analysis of pharmaceutical firms (2018)

DMU	*Discretionary Inputs* (Capital Assets)	*Discretionary Inputs* (Total Assets)	*Non-Discretionary Inputs* (Foreign Exchange Expenditure)	*Desirable Outputs (Sales)*	*Desirable Outputs (Market Value)*	*Undesirable Outputs* (Total Liabilities)
DMU 1	1860000	8361009	237	4892461	8745720	4440730
DMU 2	3300000	7531821	351	789436	957710	909128
DMU 3	525000	3422666	58	1706454	3849421	1749912
DMU 4	120000	339694	17	332050	732359	105842
DMU 5	2788000	9157887	134	2518880	2993618	2108775
DMU 6	750000	4549927	108	2238410	4460310	2221936
DMU 7	450000	4527811	233	4371785	3442856	2978778
DMU 8	829400	3424418	105	2064340	1674409	1707542
DMU 9	1134000	5638995	92	2082058	4405091	2729396
DMU 10	59300	5011500	149	3171089	6113332	3271090
DMU 11	2504000	6473433	176	2068463	7983478	5432753
DMU 12	400000	4994239	159	3228158	1802180	3681294
DMU 13	2675000	6459012	72	1142391	1583449	1060282
DMU 14	540000	25535583	136	2092988	1398951	1424337
DMU 15	699840	1941354	43	618942	2129613	1216999
DMU 16	800000	4121000	123	2600454	8638344	2537010
DMU 17	850000	2172906	38	924547	1636055	604361
DMU 18	255000	1711380	70	1155346	930916	1245460
DMU 19	625000	6774760	262	4898861	3652050	4867173
DMU 20	576000	2599529	48	1335990	2254406	1116399
DMU 21	500000	3346911	108	2090302	2417395	2219484

4.1. Cost efficiency and congestion results

The proposed two-stage DEA framework was applied to evaluate cost efficiency and diagnose cost congestion. The first-stage results indicate variation in cost efficiency across firms, with several firms operating on the cost-efficient frontier. The second-stage analysis identifies firms 2, 5, and 13 as exhibiting cost congestion relative to the cost-efficient benchmarks. This finding suggests that excessive utilization of discretionary inputs contributes to cost

[DOI: 10.71885/ijorlu-2026-3-733]

inefficiency. In such cases, reducing discretionary inputs could improve cost efficiency without reducing desirable outputs. In contrast, other inefficient firms do not exhibit cost congestion, indicating that their inefficiency may be attributed to operational or managerial factors rather than excessive input utilization. These findings highlight the importance of distinguishing cost congestion from standard cost inefficiency. Improving cost efficiency and reducing congestion can contribute to more sustainable resource utilization and enhance the economic and environmental performance of pharmaceutical firms. The detailed results are reported in Table 3.

Table 3. Cost efficiency and cost congestion results.

DMU	Cost efficiency	Cost congestion
DMU1	1.000	0
DMU2	0.111	3.446×10^5
DMU3	1.000	0
DMU4	1.000	0
DMU5	0.246	3.536×10^5
DMU6	0.708	0
DMU7	1.000	0
DMU8	0.584	0
DMU9	1.000	0
DMU10	1.000	0
DMU11	1.000	0
DMU12	1.000	0
DMU13	0.160	4.986×10^4
DMU14	0.727	0
DMU15	1.000	0
DMU16	1.000	0
DMU17	1.000	0
DMU18	0.990	0
DMU19	1.000	0
DMU20	1.000	0
DMU21	0.875	0

Figure 1 illustrates the relationship between cost efficiency and cost congestion across the evaluated pharmaceutical firms. The majority of DMUs exhibit zero congestion regardless of their efficiency level, indicating that their inefficiency is primarily managerial rather than structural. However, DMU2, DMU5, and DMU13 clearly show positive congestion values and low efficiency scores. This pattern confirms that their cost inefficiency is associated with excessive utilization of discretionary inputs. The visual evidence supports the second-stage diagnostic results reported in Table 3.

[DOI: 10.71885/ijorlu-2026-3-733]

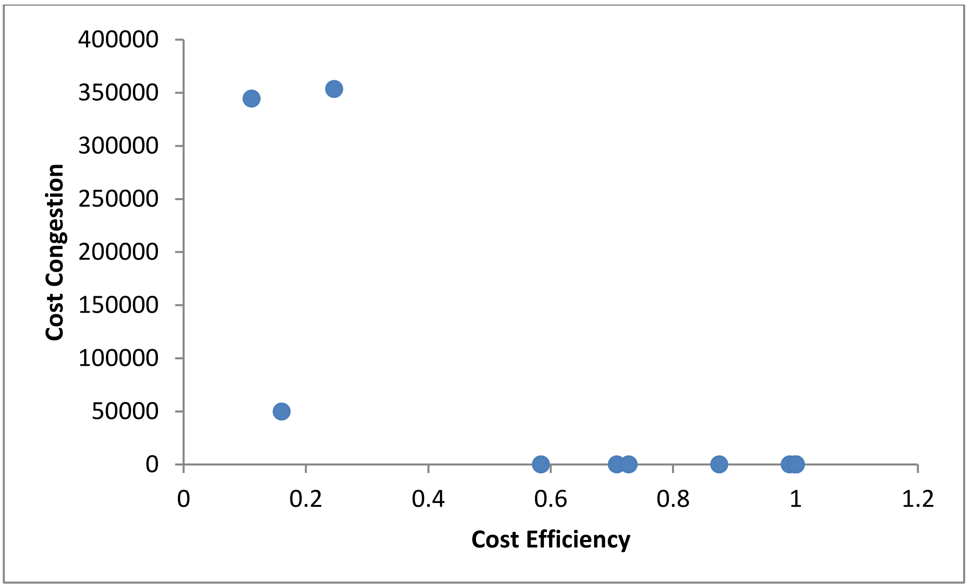

Fig. 1 Cost congestion vs cost efficiency

4.2. Discussion

The empirical results provide important insights into the relationship between cost efficiency, resource utilization, and sustainable operational performance in pharmaceutical firms. Firms identified as cost-efficient demonstrate an effective allocation of financial resources, indicating their ability to generate desirable economic outcomes without excessive input consumption. From a sustainability perspective, efficient resource utilization is closely associated with reduced waste of financial and physical resources, which may indirectly contribute to lower environmental pressure through improved operational discipline and reduced unnecessary production activities. In contrast, firms exhibiting cost congestion, such as DMUs 2, 5, and 13, show evidence of excessive utilization of discretionary financial resources relative to their output levels. This pattern suggests the presence of structural inefficiencies, where additional resource allocation does not translate into proportional performance gains. Such inefficiencies may lead to overinvestment in physical infrastructure, excessive capacity, or inefficient operational processes, which can increase energy consumption, material use, and associated environmental impacts. Therefore, cost congestion not only reflects economic inefficiency but may also indicate unsustainable resource utilization patterns. These findings highlight the importance of integrating cost efficiency analysis into broader sustainability assessment frameworks. Improving cost efficiency and eliminating congestion can enhance both economic performance and resource sustainability by encouraging firms to optimize their use of financial and operational resources. In the pharmaceutical industry, where production processes can be resource-intensive, efficient resource management plays a critical role in reducing environmental burden while maintaining economic competitiveness. Overall, the proposed framework provides a useful tool for identifying inefficient and congested firms and offers valuable information for managers and policymakers seeking to promote sustainable and efficient resource utilization in the pharmaceutical sector.

[DOI: 10.71885/ijorlu-2026-3-733]

4.3. Managerial implications

The findings of this study provide several important implications for managers, policymakers, and environmental planners. For managers, the proposed model helps identify whether inefficiency is caused by poor management practices or excessive resource utilization. In the case of cost congestion, increasing investment or expanding resource allocation may worsen performance rather than improve it. Therefore, managers should focus on optimizing existing resource utilization rather than increasing input levels. For policymakers, the model provides a useful tool for monitoring resource utilization efficiency in the pharmaceutical sector. Identifying firms with excessive resource consumption can help policymakers design appropriate regulatory and policy interventions to promote more efficient and sustainable industrial operations. From an environmental perspective, improving cost efficiency and eliminating congestion can contribute to more sustainable resource use. Efficient resource allocation reduces unnecessary production capacity, energy consumption, and material use, thereby supporting environmental sustainability objectives.

5. Conclusion

This study demonstrates that relying solely on conventional cost-efficiency measures can obscure the presence of cost congestion in inefficient decision-making units (DMUs). To overcome this limitation, a novel two-stage integrated DEA framework was developed that combines cost-efficiency assessment with a systematic diagnostic procedure for identifying and quantifying cost congestion. By simultaneously considering discretionary and non-discretionary inputs as well as desirable and undesirable outputs, the proposed model provides a more realistic representation of production processes under heterogeneous and real-world conditions.

The empirical application to 21 Iranian pharmaceutical companies reveals important insights. While several firms operate on the cost-efficient frontier, a subset of inefficient DMUs (notably DMUs 2, 5, and 13) exhibits significant cost congestion. This finding underscores a critical managerial implication: cost inefficiency does not always stem from poor operational management; in some cases, it results from structural congestion caused by excessive utilization of discretionary inputs. In such situations, further expansion of resources may exacerbate rather than alleviate performance problems.

The primary contribution of this research lies in advancing DEA methodology from traditional descriptive benchmarking toward actionable, prescriptive diagnosis. By explicitly distinguishing managerial inefficiency from structural cost congestion relative to cost-efficient benchmarks, the framework equips decision-makers with clearer guidance for implementing targeted interventions—such as process redesign, resource reallocation, or capacity optimization—rather than indiscriminate input increases.

From a practical perspective, the proposed approach offers a robust decision-support tool for managers and policymakers in resource-intensive industries like pharmaceuticals. In the Iranian context, where firms face challenges such as currency fluctuations, regulatory constraints, and high operational costs, identifying congested units can lead to more sustainable resource utilization, reduced financial risks, and improved environmental performance by minimizing unnecessary production capacity and waste.

Future research could extend this framework in several promising directions, including dynamic and multi-period analysis, integration with stochastic or fuzzy DEA to handle

[DOI: 10.71885/ijorlu-2026-3-733]

uncertainty, incorporation of network structures for multi-stage production processes, or combination with machine learning techniques for large-scale applications. Such extensions would further enhance the model's applicability to sustainability-oriented performance management and proactive policy design.

In conclusion, the integration of cost efficiency and cost congestion analysis within a unified DEA framework represents a meaningful step toward more nuanced and effective performance evaluation. This study not only enriches the theoretical literature on DEA but also provides practical value for achieving economic efficiency and environmental sustainability in complex production systems.

References

1. Charnes, A., Cooper, W. W., & Rhodes, E. (1978) Measuring the efficiency of decision-making units. European Journal of Operational Research, 2(6), 429–444. https://doi.org/10.1016/0377-2217(78)90138-8.
2. Banker, R. D., Charnes, A., & Cooper, W. W. (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis. Management Science, 30(9), 1078–1092. https://doi.org/10.1287/mnsc.30.9.1078.
3. Emrouznejad, A., Yang, G. L., & Amin, G. R. (2019). A novel inverse DEA model with application to allocate the CO2 emissions quota to different regions in Chinese manufacturing industries. Journal of the operational research society, 70(7), 1079-1090. https://doi.org/10.1080/01605682.2018.1489344.
4. Wang, Z., & Zelenyuk, V. (2024). Data envelopment analysis: from foundations to modern advancements. Foundations and Trends in Econometrics, 13(3), 170-282. https://doi.org/10.1561/0800000040.
5. Farrell, M.J. (1957) The measurement of productive efficiency. Journal of the Royal Statistical Society: Series A (General) 120(3):253–281. https://doi.org/10.2307/2343100.
6. Färe, R., & Grosskopf, S. (1985). A nonparametric cost approach to scale efficiency. The Scandinavian Journal of Economics, 594-604. https://doi.org/10.2307/3439974.
7. Daraio, C., & Simar, L. (2021) Advanced robust and nonparametric methods in efficiency analysis: methodology and applications. Springer, New York. https://doi.org/10.1007/978-0-387-35231-2.
8. Emrouznejad, A., Marra, M., Yang, G. L., & Michali, M. (2023). Eco-efficiency considering NetZero and Data Envelopment Analysis: A critical literature review. IMA Journal of Management Mathematics. https://doi.org/10.1093/imaman/dpad002.
9. Khezri, S., Dehnokhalaji, A., & Hosseinzadeh Lotfi, F. (2021). A full investigation of the directional congestion in data envelopment analysis. RAIRO-operations research, 55, S571-S591. https://doi.org/10.1051/ro/2019092.
10. Tone, K., & Sahoo, B. K. (2004). Degree of scale economies and congestion: A unified DEA approach. European journal of operational research, 158(3), 755-772. https://doi.org/10.1016/S0377-2217(03)00370-9.
11. Banker, R. D., & Morey, R. C. (1986) Efficiency analysis for exogenously fixed inputs and outputs. Operations Research, 34(4), 513–521. https://doi.org/10.1287/opre.34.4.513.
12. Seiford, L. M., & Zhu, J. (2002) Modeling undesirable factors in efficiency evaluation. European Journal of Operational Research, 142(1), 16–20. https://doi.org/10.1016/S0377-2217(01)00293-4.
13. Kapłan, R. (2026). Understanding DEA Efficiency Results in Energy Technologies: The Role of Non-Discretionary Inputs and Undesirable Outputs. Energies, 19(10), 2417. https://doi.org/10.3390/en19102417.
14. Khodadadipour, M., Hadi-Vencheh, A., Behzadi, M. H., & Rostamy-Malkhalifeh, M. (2021). Undesirable factors in stochastic DEA cross-efficiency evaluation: an application to thermal power plant energy efficiency. Economic Analysis and Policy, 69, 613-628. https://doi.org/10.1016/j.eap.2021.01.013.
15. Liu, W., et al. (2015). Two-stage DEA models with undesirable input-intermediate-outputs. Omega, 56, 74–87. https://doi.org/10.1016/j.omega.2015.03.009.
16. Moghaddas, Z., Tosarkani, B. M., & Yousefi, S. (2021). A developed data envelopment analysis model for efficient sustainable supply chain network design. Sustainability, 14(1), 262. https://doi.org/10.3390/su14010262.
17. Mergoni, A., Emrouznejad, A., & De Witte, K. (2025). Fifty years of Data Envelopment Analysis. European Journal of Operational Research, 326(3), 389–412. https://doi.org/10.1016/j.ejor.2024.12.049.
18. Panwar, A., Olfati, M., Pant, M., & Snasel, V. (2022). A Review on the 40 Years of Existence of Data Envelopment Analysis Models: Historic Development and Current Trends: A. Panwar et al. Archives of computational methods in engineering, 29(7), 5397-5426.

[DOI: 10.71885/ijorlu-2026-3-733]

19. Färe R, Svensson L (1980) Congestion of production factors. Econometrica, 48, 1745–1753. https://doi.org/10.2307/1911932.
20. Cooper, W. W., Gu, B., & Li, S. (2001). Comparisons and evaluations of alternative approaches to the treatment of congestion in DEA. European Journal of Operational Research, 132(1), 62-74. https://doi.org/10.1016/S0377-2217(00)00113-2.
21. Wei, Q., & Yan, H. (2004). Congestion and returns to scale in data envelopment analysis. European Journal of operational research, 153(3), 641-660. https://doi.org/10.1016/S0377-2217(02)00799-3.
22. Noura, A. A., Lotfi, F. H., Jahanshahloo, G. R., Rashidi, S. F., & Parker, B. R. (2010). A new method for measuring congestion in data envelopment analysis. Socio-economic planning sciences, 44(4), 240-246. https://doi.org/10.1016/j.seps.2010.06.003.
23. Tavassoli, M., Saen, R. F., & Zanjirani, D. M. (2020). Assessing sustainability of suppliers: A novel stochastic-fuzzy DEA model. Sustainable production and consumption, 21, 78-91. https://doi.org/10.1016/j.spc.2019.11.001.
24. Yousefi, S., Shabanpour, H., Ghods, K., & Saen, R. F. (2023). How to improve the future efficiency of Covid-19 treatment centers? A hybrid framework combining artificial neural network and congestion approach of data envelopment analysis. Computers & Industrial Engineering, 176, 108933. https://doi.org/10.1016/j.cie.2022.108933.

[DOI: 10.71885/ijorlu-2026-3-733]