Appendix 1:

Three-stage data envelopment analysis is a method developed based on DEA. It is mainly used to evaluate the relative efficiency of decision-making units with multiple inputs and multiple outputs, especially to reflect the efficiency of decision-making units more realistically after removing the influence of environmental factors and random disturbances.[26] The modelling approach consists of three main stages.

In the first stage, which involves conducting a conventional DEA analysis, the DEA model was introduced in 1978 by American operations researchers Charnes, Cooper, and others.[27] It is a non-parametric, non-stochastic model designed for measuring and evaluating efficiency, based on the concept of the "production frontier."[28] The model employs a linear programming approach to construct a production frontier, utilizing input and output indicators from decision-making units. Effective units are positioned on the frontier, while ineffective ones are positioned below it. This arrangement allows for the measurement of the extent to which units deviate from the frontier.[29] There are two types of DEA models: the first is the CCR model, which assumes constant returns to scale. Under this model, an increase in input will proportionately increase output, implying that the sector size does not impact production efficiency. However, this assumption often proves challenging in practice; the policy system and economic development levels may prevent maintaining sector production at a reasonable size, and thus obscure the determination of size impact on production efficiency. Based on these limitations, Charnes and other scholars revised the CCR model and proposed the BCC model, which assumes variable returns to scale.[30]

In addition, DEA models can be categorised as input-oriented or output-oriented. The former emphasizes reducing inputs while maintaining constant outputs, whereas the latter focuses on increasing outputs while keeping inputs constant.[31,32] Considering that the returns to scale for health resource allocation are variable, this study employs the input-oriented BCC model. The model equations are presented below:

$$\min\theta - \varepsilon(\hat{e}^{T}S^{-} + e^{T}S^{+})$$
$$\text{s.t.}\begin{cases} \sum_{j=i}^{n} X_j\lambda_j + S^{-} = \theta X_0 \\ \sum_{j=i}^{n} Y_j\lambda_j + S^{+} = Y_0 \\ \lambda_j \geq 0, S^{-}, S^{+} \geq 0 \end{cases} \quad (1)$$

Where, j=1,2,...,n denote decision units, X and Y are input and output vectors, respectively.

The efficiency value measured by the BCC model is called the combined Technical Efficiency (TE), and it can be further decomposed into the product of Scale Efficiency (SE) and Pure Technical Efficiency (PTE), i.e., TE=SE*PTE.[33,34]

TE measures the ability of a decision-making unit to optimise output with specific inputs under fixed production conditions and provides a comprehensive assessment of resource allocation and use efficiency. PTE reflects the impact of management skills and technical expertise on production efficiency, helping to evaluate whether a decision-making unit's management and technology are optimal. SE assesses the impact of production scale on a decision-making unit's efficiency, focusing on whether the scale is optimised.

The second stage typically involves constructing a regression model akin to Stochastic Frontier Analysis (SFA). This model estimates the influence of environmental

factors on efficiency scores through regression analysis, using environmental factors as independent variables and efficiency scores obtained in the first stage as dependent variables.[35] The SFA regression model is applied to decompose the slack variables identified in the first stage into three components: random factors, environmental factors, and managerial inefficiency.[36,37] Initially, the first-stage DEA model is analyzed to obtain the slack variables for each decision-making unit. The formula is presented as follows:

$$S_{ni} = x_{ni} - x_{ni}^{*}(n = 1,2,\dots,N; i = 1,2,\dots,I) \quad (2)$$

S_{ni} represents the slack variable for the nth input indicator of the ith decision unit, x_{ni} represents the actual value of the input indicator of each decision unit, and x_{ni}^{*} represents the predicted value of the input indicator of each decision unit. The SFA regression function is constructed using the slack variables as the response variables and the environmental factor variables as the independent variables in the analysis. The function is detailed below:

$$S_{ni} = f(Z_i; \beta_n) + v_{ni} + \mu_{ni}; i = 1,2,\dots,I; n = 1,2,\dots,N \quad (3)$$

In this function, Z_i represents the total number of environmental variables, and β_n is the value of the coefficient measured by the environmental variables. In addition, the function contains a mixed error term, $v_{ni} + \mu_{ni}$, where v_{ni} represents random error and μ_{ni} represents management inefficiency.

The SFA regression model adjusts for environmental and stochastic factors to normalize the overall technical efficiency across all decision-making units, ensuring uniform environmental conditions and stochastic influences. The mathematical expression for the function, which relates to the adjusted input variables, is presented below:

$$X_{ni}^{A} = X_{ni} + [max(f(Z_i; \widehat{\beta}_n)) - f(Z_i; \widehat{\beta}_n)] + [max(v_{ni}) - v_{ni}]$$
$$i = 1,2,\dots,I; n = 1,2,\dots,N \quad (4)$$

where X_{ni}^{A} represents the adjusted input values and Xni represents the pre-adjusted input values. The expression $[\max(f(Z_i; \hat{\beta}_n)) - f(Z_i; \hat{\beta}_n)]$ is used to place all decision-making units in a consistent external environment to ensure a fair comparison of environmental factors. Meanwhile, $[\max(v_{ni}) - v_{ni}]$ serves to adjust the random errors of all decision-making units to the same level to accurately assess their efficiency. To effectively eliminate the effects of random errors on the slack variables, further decomposition of these errors and efficiency residuals is necessary. This approach allows us to obtain the predicted random error values for each decision-making unit. For this purpose, this study employs the formula for calculating management inefficiency, as derived by Rodenyue[38], presented below:

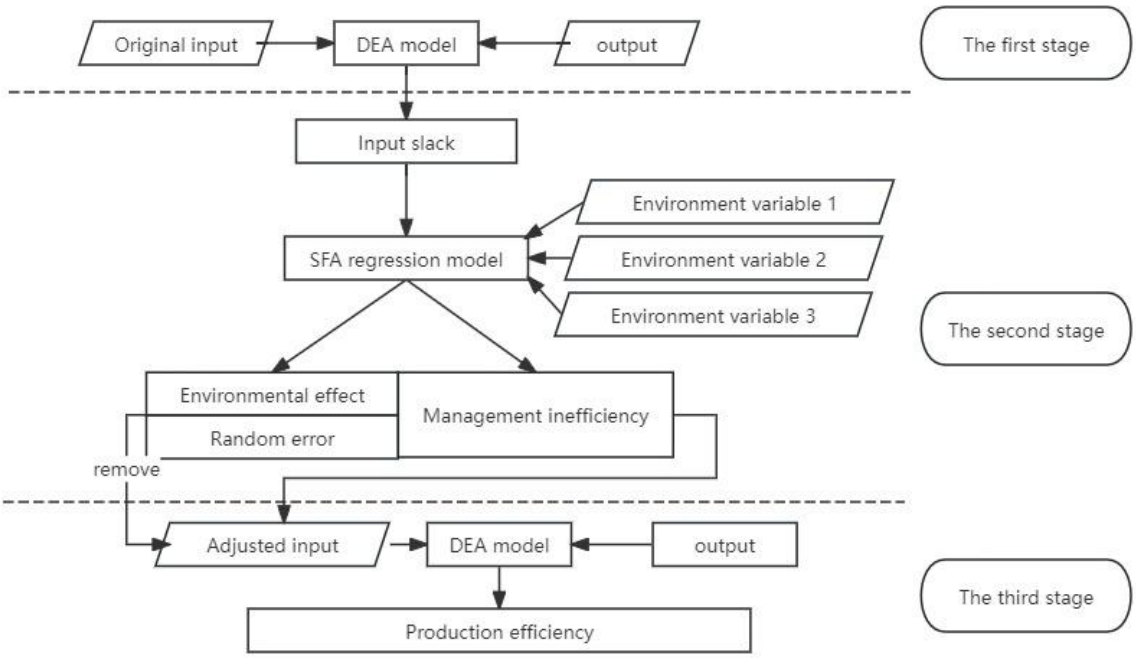
$$E(\mu|\varepsilon) = \sigma_{*}\left[\frac{\phi(\frac{\varepsilon}{\sigma})}{\Phi(\frac{\lambda\varepsilon}{\sigma})} + \frac{\lambda\varepsilon}{\sigma}\right] \quad (5)$$

Where, $\sigma_{*} = \frac{\sigma_{\mu}\sigma_{v}}{\sigma}$, $\sigma = \sqrt{\sigma_{\mu}^{2} + \sigma_{v}^{2}}$, $\lambda = \sigma_{\mu}/\sigma_{v}$, $\gamma = \frac{\sigma_{\mu}^{2}}{\sigma_{\mu}^{2}+\sigma_{v}^{2}}$.

Based on the above equation, we can derive the extent to which random error and management inefficiency factors influence the slack variable. When the value of γ variable approaches 1, it indicates a significant impact of management inefficiency; conversely, when the value of γ variable approaches 0, it indicates a significant impact of random error.

The third stage involves DEA efficiency evaluation with adjusted inputs. In this stage, efficiency evaluation is conducted using adjusted input data and original output data, after removing the influence of environmental factors and random errors. First, input indicators are adjusted based on the regression analysis results from the second stage to eliminate the interference of external environmental factors and random errors. Then, the output indicator data remain unchanged, and the adjusted input data are substituted into the

DEA model for calculation. Finally, this stage provides more accurate and realistic efficiency evaluation results, providing scientific evidence for decision-makers.



Three-stage DEA model flow chart