Improving Access to Essential Medicines in Sierra Leone via Machine Learning

Angel T-H Chung,[1] Patrick Bayoh,[3] Jatu Abdulai,[3] Lawrence Sandi,[3] Francis Smart,[4] Hamsa Bastani,[1,*] Osbert Bastani[2,*]

[1]Department of Operations, Information, and Decisions, University of Pennsylvania, Jon M. Huntsman Hall, 3730 Walnut St, Philadelphia, PA 19104, USA
[2]Department of Computer and Information Science, University of Pennsylvania, 3330 Walnut St, Philadelphia, PA 19104, USA
[3]National Medical Supplies Agency, Sierra Leone
31 Murray Town Road, Freetown, Sierra Leone
[4]Ministry of Health and Sanitation, Sierra Leone
Youyi Building Freetown, Freetown, Sierra Leone
[*]To whom correspondence should be addressed; E-mail: hamsab@wharton.upenn.edu, obastani@seas.upenn.edu.

A critical challenge in healthcare systems in Low- and Middle-Income Countries (LMICs) is the efficient and equitable allocation of scarce resources, particularly essential medicines. This problem is complicated by limited high-quality data, which restricts the applicability of traditional data-driven techniques. We propose a novel machine learning framework for essential medicines allocation, which leverages a combination of multi-task learning and decision-aware learning to improve sample efficiency and ensure equitable allocation. In collaboration with the Sierra Leone national government, our framework has been deployed nationwide as a decision support tool to help reduce waste and improve essential medicines allocation. Our evaluation using synthetic difference-in-differences analysis demonstrates a 19% increase in medicine con-

sumption, with no changes to the supply, improving access for approximately 3.7 million women and children under five. Through experimental validation, we demonstrate that our approach also significantly outperforms baseline approaches. Our work demonstrates the tangible impact of machine learning in optimizing high-stakes decisions in resource-constrained settings, improving efficiency while ensuring equity and cost-effectiveness.

Introduction

Machine learning has demonstrated enormous potential for improving healthcare, with applications ranging from screening and diagnosis (*1–3*), targeted testing (*4*), and automating medical record generation (*5*). However, success stories have largely been limited to developed nations, due to the critical need for high-quality data used to train models. Beyond the lack of quality data, the healthcare needs in developing nations also differs significantly from those in developed ones. For instance, according to a WHO health facility assessment, many African countries report less than 40% availability of maternal essential medicines (*6*). This scarcity can pose a serious threat to public health, especially for vulnerable populations that disproportionately rely on public facilities; for instance, it forces patients to either pay inflated prices in the private sector or forgo treatment altogether (*7*).

As a consequence, there is significant interest in how machine learning can be effectively leveraged to improve healthcare in global health contexts. Challenges in developing nations are often operational in nature due to infrastructure limitations such as poor stock management systems (*8*) or insufficient staff training in logistics and inventory management (*9*). For example, in 2010, Sierra Leone launched the Free Health Care Initiative (FHCI), one of the largest healthcare initiatives and a top priority in its post-civil war recovery (*10*), providing free medical care and products to pregnant women and children under five. However, these

essential medicines are only distributed to healthcare facilities once a quarter, which can lead to significant shortages in one location even if there is adequate supply in other locations. Efforts such as Project Last Mile (PLM)[1] have attempted to directly improve this supply chain, but have encountered logistical obstacles that impede progress and country-wide scalability (*11, 12*). For example, despite operating from 2018-2023, PLM's expansion of electronic logistics management systems to improve last-mile delivery reached only 15 government hospitals and 103 health facilities (*13*)—approximately 8% of total facilities. Planning ahead is also challenging, since supplies are often donated and can vary enormously every quarter.

Rather, a low-cost and promising avenue to scalably reduce shortages is to better match limited supply with patient demand using a combination of prediction and optimization (*14, 15*)—in particular, given accurate facility-level demand forecasts for a product, existing supply chain optimization techniques (*16–18*) can optimize facility-level allocations to minimize unmet patient demand. However, forecasting demand is difficult due to the lack of high-quality data, which is further exacerbated by high demand variability due to natural disasters or disease outbreaks (*7, 19, 20*). Existing strategies employed in global health contexts include relying on historical consumption data, morbidity-based forecasting, or proxy estimates; however, these approaches tend to have limited accuracy (*21–23*). Furthermore, due to the lack of modern computing infrastructure, 82% of deployments of these techniques rely either on ad hoc manual forecasting (*24*) or on complex Excel spreadsheets with error rates as high as 40% (*25*). For instance, Sierra Leone previously deployed an Excel tool developed by Crown Agents, a non-profit international development organization, to support demand forecasting and supply allocation for maternal essential medicines; however, our analysis finds that their tool produces inaccurate forecasts resulting in excessive medicine shortages.

In this paper, we introduce a novel machine learning framework to optimize constrained

[1]Project Last Mile is an intensive effort to adapt The Coca-Cola Company's highly effective supply chain strategies to improving last mile delivery and the availability of essential medicines in LMICs.

resource allocation, which is designed to satisfy two key criteria: it makes effective use of limited and noisy historical data, and it is scalable and can be deployed in limited-compute environments. At a high level, our system leverages machine learning (specifically, random forests) to predict demand based on features constructed from historical features, and then applies stochastic optimization to compute the best allocation based on this model's predictions. To tackle data scarcity and quality challenges, our system uses a multi-task learning strategy to share data across different healthcare facilities (*26*), along with a novel decision-aware learning algorithm (*27–32*) that preferentially allocates predictive power to predictions that have the greatest impact on improving the downstream allocation optimization. Furthermore, it uses catalytic priors (*33*) and auxiliary data sources to mitigate data inequity (i.e., where poorer facilities have lower-quality data and therefore noisier forecasts).

In close collaboration with the Sierra Leone national government, we deployed our system nationwide to improve allocation decisions for Sierra Leone's FHCI policy. This deployment was implemented via a staggered rollout across 1,058 government healthcare facilities across the country, providing a source of natural variation for us to perform an econometric analysis of the efficacy of our system. Our analysis finds a 19% increase in overall consumption of essential medicines with no changes to the supply, indicating a significant improvement in patient access. Importantly, our intelligent system achieved these gains while being extremely cost-effective, requiring only a $30 monthly server fee and no additional workforce. The success of our deployment highlights the ability for machine learning to address critical operational challenges in improving healthcare delivery in developing nations.

Esssential Medicines Allocation System

Maternal mortality remains one of the most pressing global health challenges, particularly in developing countries where access to essential health products is often limited by inefficient

allocation systems (*34*). In Sierra Leone, despite the FHCI providing free medical care to pregnant women and children, maternal mortality rate stands at 717 per 100,000 live births — one of the highest in the world (*35, 36*).

Lack of access to essential medicines is one of the key contributors to preventable maternal deaths (*34*). In Sierra Leone, the National Medical Supplies Agency (NMSA) (part of the Ministry of Health and Sanitation (MoHS)) was created to manage the procurement and distribution of medicines and medical supplies to public health facilities across the nation, including over 70 essential medicines for women and children. Supply of these medicines largely relies on donations from international organizations. Prior to our collaboration, the NMSA distributed these medicines across the country following a centralized two-level push system (*37*), where the NMSA first allocates to 16 districts, and then districts allocate to individual health facilities in their catchment (see details in Supplement §1.3). Allocation decisions were largely made via an Excel tool with district managers making significant manual adjustments since they perceived that the tool's estimates did not accurately reflect the actual needs of health facilities.

Despite these efforts, there remained a significant shortage of medicines across the nation. Critically, this shortage occurred even when the total supply was adequate—some health facilities received a surplus of medicines[2] while others suffered shortages. For instance, the District Health Information Software 2 (DHIS2) shows significant heterogeneity in both stockouts and wastage—e.g., in the district Tonkolili in 2022 Q1, 10% of facilities that had previous stockouts continued to face supply shortages, while 22% of facilities with available stock had excess supply that could have been redistributed. These findings suggest that more efficient allocation strategies—that better match supply and demand—are a promising path to significantly reducing shortages.

Fig. 1 illustrates the system we developed and deployed. Our system begins by pulling detailed supply and consumption data from government databases, including monthly data on

[2]Surpluses often did not carry over fully to the next quarter due to significant reported waste and expiration.

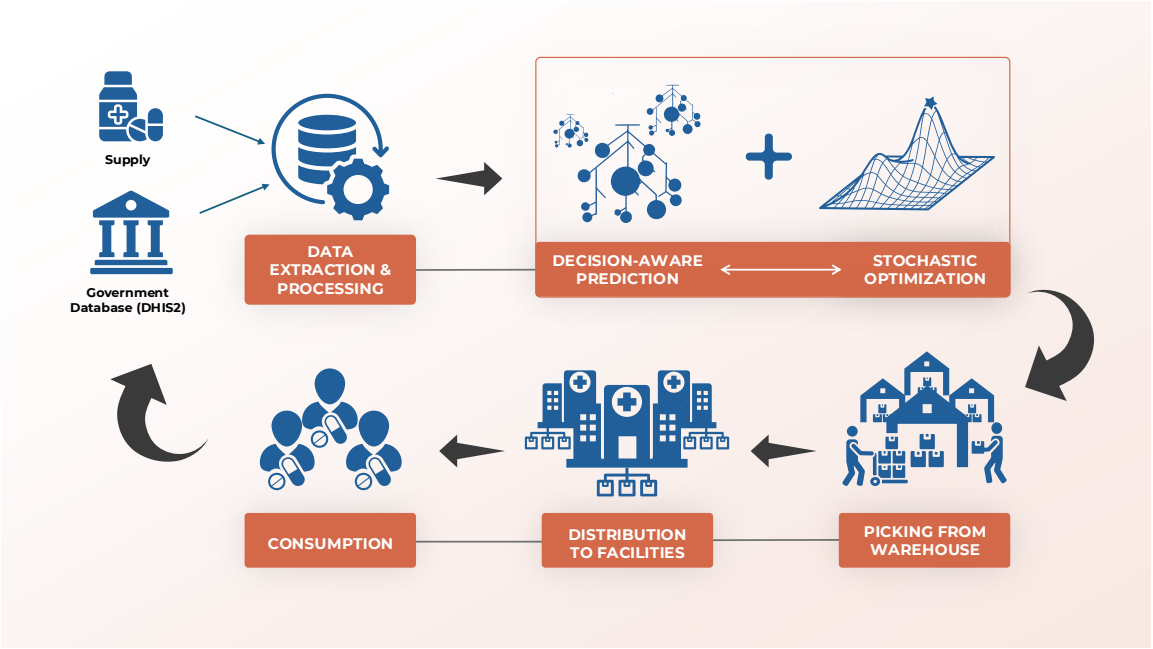

Figure 1: **System Overview.** Every quarter, the system extracts and processes data from supply records and the DHIS2 government database. It then trains a decision-aware prediction model that informs a stochastic optimization procedure to make allocation decisions. The system further provides a picking list for frontline workers to collect supplies from designated warehouses, enabling efficient distribution to local health facilities. Finally, the resulting patient consumption data is recorded for informing predictions and decisions in subsequent quarters.

amounts allocated, consumed, and remaining stock for each medical product at each health facility (Supplement §1.1). We then perform significant pre-processing to ensure data reliability and construct informative features for prediction (Supplement §1.2).

Our algorithm (the box in Fig. 1) can be divided into two components. First, we predict the demand distribution for each facility-product pair using a novel decision-aware machine learning framework described in the next section. Second, given the demand forecasts, our optimization algorithm produces allocations designed to minimize the expected shortage of medicines. In particular, for each product, it aims to minimize *unmet demand*—the total number of eligible patients turned away—across facilities. To account for the stochastic nature of demand, we use

the expected unmet demand according to the probabilistic demand forecasts. Unlike the prior two-level approach, we output allocations directly from the central stock to individual health facilities. This optimization can be solved efficiently via a linear program using a technique known as sample average approximation (*38*). Once the allocation is determined, our system assigns each batch of supplies to a warehouse based on proximity, stock availability, and product expiration dates (Supplement §1.4). Finally, the ensuing patient consumption data is recorded in government databases, and used to retrain our predictive models and optimize the next set of allocation decisions in subsequent quarters.

Machine Learning for Demand Forecasting

Next, we describe our machine learning framework for predicting the demand distribution used in our optimization algorithm. First, we construct a training dataset from historical demand data.[3] Given this dataset, we could apply a traditional strategy for time series forecasting such as ARIMA (*39*). However, these approaches work poorly due to data scarcity—on average, only 28 reliable data points are available for each product-facility pair, which is too few to train accurate forecasting models. Demand tends to be especially unpredictable in developing nations due to variability in when patients seek out healthcare resources (*19*).

Instead, we design a machine learning framework that leverages three key techniques: (i) multi-task learning to share data across facilities, (ii) catalytic priors to regularize the model in data-poor regions to mitigate data inequity, and (iii) a novel decision-aware learning algorithm to focus predictive power on facilities that are most relevant to the downstream optimization problem. We summarize our techniques below, and provide details in Supplement §1.5.

First, our system uses a multi-task learning strategy where, for each grouping of related health products, we train a single demand prediction model across all facility-product pairs. This

[3]We describe our data preprocessing pipeline to address unreliable or missing data in Supplement §1.2 & §1.5.2.

facilitates knowledge transfer from locations with more available data to ones with less available data (*26, 40, 41*). In particular, our strategy constructs features that facilitate generalization across facility-product pairs (e.g., average demand in the past year), as well as features that capture trends specific to a given facility-product pair (e.g., the facility type and location, product fixed effect). Then, we train a random forest to predict demand from these features. This strategy enables us to transfer knowledge across facilities and products while accounting for facility- and product-specific trends to the best degree possible (Supplement §1.5.1).

Second, while multi-task learning can improve performance in data-poor locations, there are still systematic differences between data-poor and data-rich locations (e.g., missing data often arises disproportionately in poorer regions due to staffing shortages). Such missing-not-at-random data (*42*) can lead to *covariate shift*, thereby reducing prediction accuracy for data-poor locations. To mitigate this data inequity, our system leverages catalytic priors (*33*) to regularize predictions for data-poor locations towards a simpler, less biased population-based model. In particular, we first use census and satellite data to estimate catchment population, and then estimate demand proportionally to the catchment population; this strategy ensures complete coverage of all facilities without biases due to low-quality data. Then, we integrate this simple model as a catalytic prior for our machine learning model, regularizing predictions for data-poor locations towards the population-based model.

Third, our system leverages decision-aware learning (*29, 30, 32, 43*) to focus predictive power on instances most relevant to the downstream optimization problem. Intuitively, not all predictions matter equally—for example, since our goal is to prevent unmet demand, we are primarily interested in how much we should stock facilities that are *likely* to be insufficiently stocked for a particular product. Whereas a standard machine learning approach would treat all facility-product pairs equally when training a prediction model, our decision-aware learning algorithm focuses attention on facility-product pairs deemed more important to the decision-

making objective—in this case, it upweights observations corresponding to likely under-stocked facilities. We found that existing decision-aware learning algorithms were either computationally intractable at our scale or incompatible with the rest of our prediction and optimization pipeline; thus, we develop a novel decision-aware learning approach, which can be easily integrated with existing data pipelines (Supplement §1.5.3). Prior to deploying our framework, we validated it on historical data by showing that it outperforms several baselines on a held-out test set, including existing decision-aware and traditional learning-based approaches (*32, 43*), a linear program developed by the World Health Organization (*44*) as well as population-based models. Our results demonstrate that our algorithm yields improvements in computational efficiency and decision quality compared to these baselines (see Table S1 in Supplement §1.5.3).

Deployment

In collaboration with the NMSA, in May 2023, we deployed our system in 5 of the 16 districts in Sierra Leone to perform facility-level allocations for the second quarter (June through August). These districts—Tonkolili, Falaba, Karene, Kono, and Pujehun (see map of treated facilities in Fig. 2)—were selected by the central government based on a randomized allocation schedule. Prior to our deployment, the government had already established both the overall supply as well as the total amounts to be allocated to control districts. The remaining supply was then to be allocated to the treatment districts, maintaining the independence of supply quantities between the two groups. Additional deployment details are in Supplement §2.1.

Beyond the performance of the allocation system, a critical priority was ensuring that it could be seamlessly integrated into NMSA's existing workflows. To this end, we conducted two targeted training sessions for policymakers and frontline workers, providing the necessary technical knowledge for operating our tool and understanding its implications. Implementation emphasized stakeholder engagement at every level—outputs were presented in a familiar format that matched

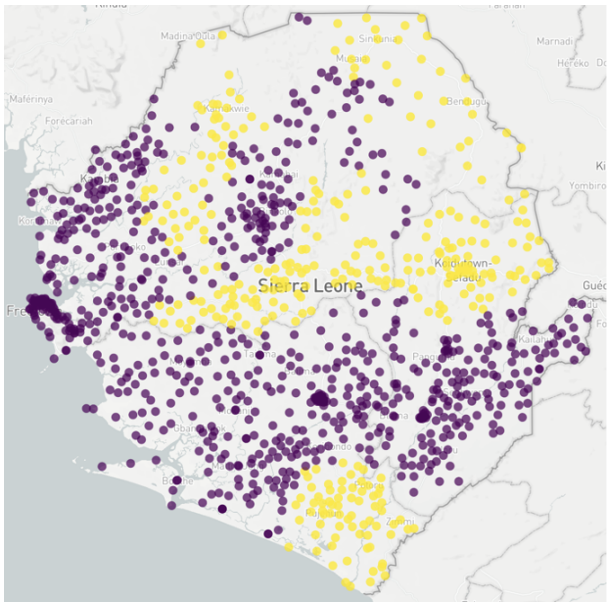

Figure 2: **Map of treatment distribution.** Yellow dots denote treated facilities (i.e., those in the Tonkolili, Falaba, Karene, Kono, and Pujehun districts), and purple dots denote control facilities (i.e., those in the Kailahun, Kenema, Bombali, Koinadugu, Kambia, Port Loko, Bo, Bonthe, Moyamba, Western Area Rural, and Western Area Urban districts).

pre-existing workflows, and were reviewed by NMSA management as well as district pharmaceutical managers prior to finalization. This process facilitated a smooth transition to the new allocation system while ensuring alignment with national and local healthcare priorities. Despite variability in communication efficiency and workforce capacity across resource-constrained districts, our system achieved a high compliance rate—measured as the absolute difference between the actual and algorithmic allocations—ranging from 89% to 100% (see Table S2 in Supplement §2.3). Given the high compliance rates, stakeholder buy-in, and early indications of improved efficiency (detailed in the following section), the national government adopted our system and expanded its use to all public-sector health facilities in Sierra Leone beginning in the third quarter of 2023, and it continues to be used throughout the country today.

Evaluation

We evaluate the effectiveness of our decision support tool at improving allocation efficiency. While our optimization objective aims to minimize unmet demand, we do not directly observe this quantity (we only observe when patients receive medicine, not when they are turned away due to stockouts). Thus, in our evaluation, we examine the equivalent objective of maximizing *patient consumption*, which is directly observed—since patient demand (which is fixed but unobserved) equals consumption plus unmet demand, maximizing total consumption is equivalent to minimizing total unmet demand. We provide an overview of our analyses here, and provide details in Supplement §2; our results are summarized in Figs. 3 & 4.

Our main analysis focuses on 2023 Q2, where our system was deployed in a randomly selected subset of districts, providing natural variation enabling us to estimate causal effects. In particular, we estimate how much our system changed patient consumption levels in treated districts (i.e., districts where our system was deployed) compared to what would have happened without our intervention, known as the Average Treatment Effect on the Treated (ATT). We estimate the ATT using a balanced panel dataset of 312 facilities in treated districts and 746 facilities in control districts using time series data beginning in 2022 Q3 (when the prior Excel allocation tool was implemented)[4] through 2023 Q3 (after which our tool was used nationwide). Given the limited number of treated districts, we use a Synthetic Difference-in-Differences (SynthDiD) regression (*45*)—which exploits temporal variation to improve statistical power—to analyze the impact of our deployment.[5] We provide details on our regression specification in Supplement §2.2.

[4]Prior to the implementation of the Excel tool by Crown Agents, allocation procedures were highly inconsistent, rendering the data unreliable.

[5]SynthDiD combines the strengths of Difference-in-Differences (*46*) (which assumes similar trends between the treated and untreated groups in the pre-treated period) and Synthetic Controls (*47*) (which constructs a synthetic control group that is similar to the treated units in terms of observed characteristics and outcome trends in the pre-treated period). SynthDiD ensures that differences between treated units and synthetic control units remain stable prior to treatment.

Results for our main analysis are shown in Fig. 3 (time series trends for the treatment and synthetic control groups), and the "SynthDiD" row in Fig. 4 (average improvement)—in particular, the five treated districts experienced a statistically significant increase of 19% ($p < 0.01$) in consumption.[6] These results demonstrate that our deployment substantially improves consumption. We validate our SynthDiD approach using a standard event-study analysis (*48*), which shows that there are no statistically significant differences between treated and control units prior to our intervention and that the change in consumption emerges only after our system was deployed; see Supplement §2.2.

Next, to study the impact of compliance on the effectiveness of our approach, we use instrumental variables (IVs) to estimate the effect of the treatment on compliant districts; we provide details on our analysis strategy in Supplement §2.3, and detailed results in Table S3 of that section. Districts with higher compliance to our algorithmic allocations (i.e., >95%, including Tonkolili, Falaba, and Karene) showed even stronger effects, with a 36% increase in consumption ($p < 0.001$). This is notable because, after the government rolled out our system to the entire country in Q3, all districts reported full compliance.

To understand the equity implications of our deployment, we examined previously under-served facilities (i.e., facilities that experienced at least one stockout in the data prior to our deployment). We found that these facilities saw an even more pronounced increase in consumption of 44% ($p < 0.01$) ("Under-Served" row of Figure 4; details in Supplement §2.4), suggesting that our system successfully addresses potential biases from uneven data quality and availability, leading to more equitable resource allocation. Overall, these results suggest that our system has substantially improved access to essential healthcare resources in treated districts.

As a robustness check, we perform a standard DiD analysis instead of using SynthDiD; we find consistent results, with a 21% increase in consumption ($p < 0.01$) ("DiD" row of

[6]The % increase is calculated from SynthDiD counterfactual estimates as: (Average treatment outcome − Average counterfactual outcome)/(Average counterfactual outcome).

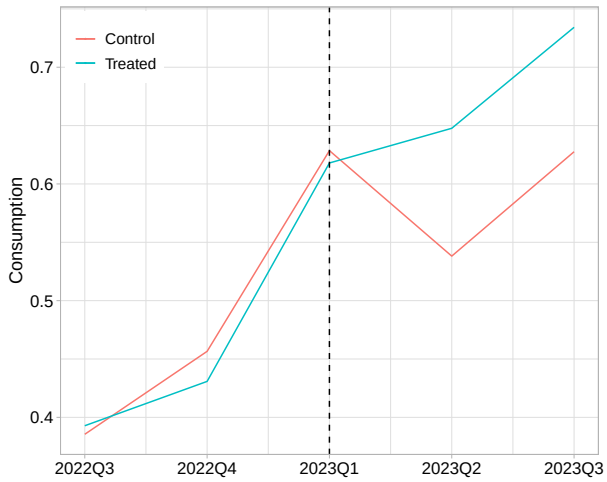

Figure 3: **2023 Q2 Deployment Result.** ATT from SynthDiD: This plot shows the estimated ATT of our 2023 Q2 deployment using SynthDiD. The x-axis shows time in quarters, and the y-axis shows normalized consumption. The solid lines indicate the consumption pattern of the treated (green) and control (red) groups. The vertical dashed line at 2023 Q1 marks the time of deployment.

Figure 4; details in Supplement §2.4). In addition, a potential concern is due to missing data from facilities that failed to record consumption in a particular month. Our main analysis drops observations with missing outcomes, but we also perform robustness checks using multiple standard imputation strategies—based on low-rank matrix completion (*49*), population-based methods, and historical average consumption. The resulting SynthDiD ATT estimates are all statistically significant and consistent with our main analysis ("Imputation" rows in Fig. 4; details in Supplement §2.4).

To further support our results, we consider an alternative analysis where we compare results across products instead of districts. In particular, we consider 25 other products that were concurrently allocated using a different, pre-existing mechanism.[7] The consumption levels for these products can be used as a control group throughout our study period across *all* districts

[7]This was either because the allocation mechanism for these products was controlled by an external third party, or because there was too little historical data to use our system.

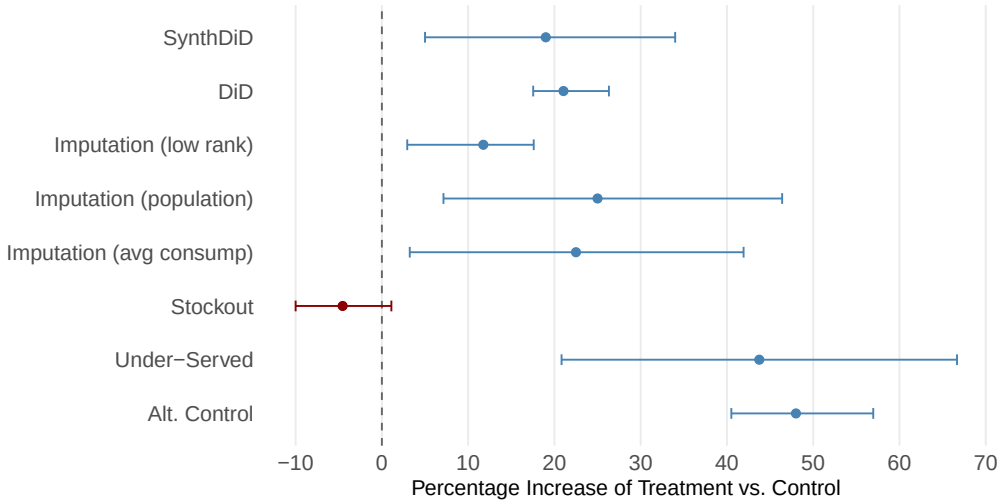

Figure 4: **ATT Results.** This figure shows ATT estimates with 95% confidence intervals for various estimation methods and robustness checks, including our main SynthDiD regression, the DiD regression, three different imputation strategies, restriction to under-served districts, and approaches using alternative controls and staggered rollout (see details in S2.4). The x-axis is percentage increase in treatment compared to control, with the dashed vertical line at 0 representing the null hypothesis of no effect. For consumption, positive treatment effect is better (blue), and for stockouts, negative treatment effect is better (red). With the exception of DiD, all methods are based on SynthDiD. As can be seen, all results are statistically significant, except for our robustness check using stockout as the outcome, which is improved but not significantly so; this is expected since it is not our primary objective.

in the country using a staggered treatment—i.e., an advantage of this analysis is that it can be performed not just for the partial deployment in 2023 Q2, but also for the nationwide implementation starting in Q3. We again find a statistically significant increase in consumption of 48% ($p < 0.01$) ("Alt. Control" row in Fig. 4; details in Supplement §2.4). In other words, analysis with this alternative control group also supports our finding that our system significantly improved allocation efficiency.

Finally, we examine the impact of our system on stockouts. While reducing stockouts might seem like a natural objective, optimizing for fewer stockouts produces highly undesirable allocations—e.g., an optimal strategy is to allocate zero supply to a small number of high-volume facilities, thereby ensuring that the remaining facilities are well-stocked. We find a directional

reduction in stockouts but it is not statistically significant ($p \approx 0.11$) ("Stockouts" row in Fig. 4; details in Supplement §2.4)—i.e., our system does not inadvertently increase stockouts.

Conclusion

Our findings provide strong field evidence of the effectiveness of our novel machine learning framework for resource allocation, significantly and equitably improving access to essential medicines in a highly-constrained environment like Sierra Leone. By replacing paper-based manual processes with automated, data-driven decision making, our system streamlines supply chain management, reduces administrative burdens, and adapts to real-time changing consumption patterns, enabling facilities to better align limited supply with patient demand.

To ensure sustainable impact, we developed a web application with an intuitive user interface, which is now owned by the Sierra Leone government. This system integrates with their government databases to automate the entire process, from data extraction and processing to generating final allocation results. Notably, it operates without any additional workforce requirements and costs only $30 per month in server fees. Their officials and staff continue to use this system at present for making allocations throughout the country.

Beyond the specific context of Sierra Leone, this work highlights the promise of machine learning to improve resource allocation in highly budget-constrained settings. While different governance structures would require different deployment processes, our framework is flexible and light-touch for easy adoption in other contexts.

References

1. J. De Fauw, *et al.*, *Nature medicine* **24**, 1342 (2018).

2. A. Esteva, *et al.*, *Nature medicine* **25**, 24 (2019).

3. K. Y. Ngiam, W. Khor, *The Lancet Oncology* **20**, e262 (2019).

4. H. Bastani, *et al.*, *Nature* **599**, 108 (2021).

5. X. Yang, *et al.*, *NPJ digital medicine* **5**, 194 (2022).

6. World Health Organization, Service availability and readiness assessment (SARA): An annual monitoring system for service delivery, *Tech. rep.*, World Health Organization (2018).

7. A. Yenet, G. Nibret, B. A. Tegegne, *ClinicoEconomics and Outcomes Research* pp. 443–458 (2023).

8. S. M. Zuma, L. M. Modiba, *World J Pharm Res* **8**, 1532 (2019).

9. C. Irene, E. A. Komomo, O. O. Augustina, E. O. Asuquo, P. O. Agada, *J Health Med Nurs* **2016**, 45 (2005).

10. J. Donnelly, *The Lancet* **377**, 1393 (2011).

11. E. Linnander, *et al.*, *PloS one* **12**, e0186832 (2017).

12. E. Linnander, K. LaMonaca, M. A. Brault, M. Vyavahare, L. Curry, *International Journal of Multiple Research Approaches* **10**, 136 (2018).

13. Project Last Mile, 2023 annual report, *Tech. rep.*, Project Last Mile (2023).

14. X. Zhu, A. Ninh, H. Zhao, Z. Liu, *Production and Operations Management* **30**, 3231 (2021).

15. F. Mbonyinshuti, J. Nkurunziza, J. Niyobuhungiro, E. Kayitare, *Processes* **10**, 26 (2021).

16. D. Bertsimas, A. Thiele, *Operations research* **54**, 150 (2006).

17. C. Yang, Z. Hu, S. X. Zhou, *Management science* **67**, 185 (2021).

18. Y. Aviv, *Management science* **47**, 1326 (2001).

19. UNICEF, *et al.*, *The State of the World's Children 2015: Reimagine the Future-Innovation for Every Child (Executive Summary)* (UN, 2015).

20. P. Yadav, *Health systems & reform* **1**, 142 (2015).

21. P. FUND, *et al.* (2016).

22. USAID — DELIVER PROJECT, Task Order 4, *Quantification of Health Commodities: A Guide to Forecasting and Supply Planning for Procurement*, Arlington, Va (2014).

23. Reproductive and Maternal Health Services Unit, Ministry of Health, *National Guidelines for Quantification, Procurement, and Pipeline Monitoring for Family Planning Commodities in Kenya*, Ministry of Health, Nairobi, Kenya (2016).

24. M. W. Fredrick, W. Muturi, *IOSR J Business Manag* **18**, 63 (2016).

25. G. Merkuryeva, A. Valberga, A. Smirnov, *Procedia Computer Science* **149**, 3 (2019).

26. R. Caruana, *Machine learning* **28**, 41 (1997).

27. P. Donti, B. Amos, J. Z. Kolter, *Advances in neural information processing systems* **30** (2017).

28. A. N. Elmachtoub, P. Grigas, *Management Science* **68**, 9 (2022).

29. D. Bertsimas, N. Kallus, *Management Science* **66**, 1025 (2020).

30. B. Wilder, B. Dilkina, M. Tambe, *Proceedings of the AAAI Conference on Artificial Intelligence* (2019), vol. 33, pp. 1658–1665.

31. K. Wang, B. Wilder, A. Perrault, M. Tambe, *Advances in Neural Information Processing Systems* **33**, 9586 (2020).

32. S. Shah, K. Wang, B. Wilder, A. Perrault, M. Tambe, *Advances in Neural Information Processing Systems* **35**, 1320 (2022).

33. D. Huang, N. Stein, D. B. Rubin, S. Kou, *Proceedings of the National Academy of Sciences* **117**, 12004 (2020).

34. J. P. Souza, *et al.*, *The Lancet Global Health* **12**, e306 (2024).

35. World Health Organization, Maternal mortality: Fact sheet (2024).

36. Y. Shafiq, *et al.*, *BMJ open* **14**, e076256 (2024).

37. C. Terwiesch, G. Cachon, *Matching supply with demand: an introduction to operations management* (McGraw-Hill, 2006).

38. J. R. Birge, F. Louveaux, *Introduction to stochastic programming* (Springer Science & Business Media, 2011).

39. K. Gilbert, *Management Science* **51**, 305 (2005).

40. H. Bastani, *Management Science* **67**, 2964 (2021).

41. K. Xu, H. Bastani, *Management Science* (2025).

42. D. B. Rubin, *Biometrika* **63**, 581 (1976).

43. N. Kallus, X. Mao, *Management Science* (2022).

44. W. H. O. W. Team, Allocation logic and algorithm to support allocation of vaccines secured through the covax facility, *Tech. rep.*, World Health Organization (2021).

45. D. Arkhangelsky, S. Athey, D. A. Hirshberg, G. W. Imbens, S. Wager, *American Economic Review* **111**, 4088 (2021).

46. D. Card, A. B. Krueger, Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania (1993).

47. A. Abadie, A. Diamond, J. Hainmueller, *Journal of the American statistical Association* **105**, 493 (2010).

48. D. Clarke, D. Pailañir, S. Athey, G. Imbens, Synthetic difference in differences estimation (2023).

49. E. Candes, B. Recht, *Communications of the ACM* **55**, 111 (2012).

50. Sierra Leone Ministry of Health and Sanitation (MoHS), Center for International Earth Science Information Network (CIESIN), Columbia University, Sierra leone national health facilities dataset version 01 (2023).

51. D. W. Martin, *et al.*, *Health security* **18**, S (2020).

52. mSupply (2024). Https://www.msupply.org.nz.

53. WorldPop, Worldpop global project population data: Estimated residential population per 100x100m grid square, https://www.worldpop.org (2021).

54. D. Weiss, *et al.*, *Nature medicine* **26**, 1835 (2020).

55. K. Didan, MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006 [Data set], https://doi.org/10.5067/MODIS/MOD13Q1.006 (2015).

56. O. Besbes, A. Muharremoglu, *Management Science* **59**, 1407 (2013).

57. W. H. Greene, *International edition, New Jersey: Prentice Hall* pp. 201–215 (2000).

58. G. A. Stevens, *et al.*, *The Lancet* **388**, e19 (2016).

59. Statistics Sierra Leone, 2015 population and housing census, *Census report*, Statistics Sierra Leone, Freetown, Sierra Leone (2015).

60. Humanitarian Data Exchange (HDX), HDX HAPI Population Dataset (2023).

61. J. Kotary, F. Fioretto, P. Van Hentenryck, B. Wilder, *arXiv preprint arXiv:2103.16378* (2021).

62. A. Agrawal, *et al.*, *Advances in neural information processing systems* **32** (2019).

63. J. Angrist, G. Imbens, Identification and estimation of local average treatment effects (1995).

64. O. C. Ashenfelter, D. Card, Using the longitudinal structure of earnings to estimate the effect of training programs (1984).

Acknowledgments

We are grateful for the close partnership offered by the Sierra Leone Ministry of Health and Sanitation and the National Medical Supplies Agency; this partnership stemmed from an earlier collaboration with Macro-Eyes and its team members (Vahid Rostami, Ashley Schmidt, Musa Komeh, Lydia Bernard-Jones, and Rene Ishiwe). We acknowledge invaluable research assistance from our team of RAs (Allan Zhang, Cheng-Ying Wu, Norris Chen, and Hingis Chang). This draft benefited from helpful feedback from participants at the Machine Learning for Health, Symposium on Artificial Intelligence in Learning Health Systems (SAIL), INFORMS Annual Conference, Marketplace Innovation Workshop, MSOM Healthcare SIG, Purdue Operations Conference, and Workshop on AI & Analytics for Social Good.

Supplementary Materials

Angel T-H Chung, Patrick Bayoh, Jatu Abadulai, Lawrence Sandi, Francis Smart,

Hamsa Bastani*, Osbert Bastani*

*Corresponding authors. Email: hamsab@wharton.upenn.edu, obastani@seas.upenn.edu

This PDF file includes:

Other Supplementary Materials for this manuscript:

Supplementary Materials

§1 describes the data and methods supporting the design of our allocation system; §2 describes the national deployment of our system and the empirical evaluation of its effectiveness.

1 Materials and Methods

We first outline the data sources (§1.1) and detail how raw data are processed and constructed for our machine learning model and evaluation (§1.2). Second, we review the allocation mechanism used in Sierra Leone prior to adopting our approach (§1.3). Then, we formalize resource allocation as an optimization problem (§1.4). Finally, we present our end-to-end machine learning pipeline for demand estimation and decision-aware allocation (§1.5), which spans:

- Multi-Task Learning (§1.5.1): we exploit cross-facility and cross-product patterns to improve predictive accuracy with limited data.
- Catalytic Priors (§1.5.2): we construct catalytic priors from population estimates to address data quality issues such as censoring and missing values.
- Decision-Aware Learning (§1.5.3): we propose a novel method for re-weighting training data to align the prediction objective with the downstream optimization objective, thus shifting the focus from prediction accuracy to improving public health.

1.1 Datasets

List of public health facilities. We obtain information on the ID, latitude, longitude, and type of public health facilities in Sierra Leone (*50*); this data was cross-verified with frontline staff at the National Medical Supplies Agency (NMSA).

Consumption data. To construct outcomes and features for demand prediction, we extract data from the District Health Information Software 2 (DHIS2) used by Sierra Leone Ministry of Health and Sanitation (MoHS) to collect and manage health data. The country transitioned from paper-based to electronic reporting of health data in public health facilities in 2019 (*51*). We extract monthly facility-level data on consumption, opening balance, closing balance, and stockouts of 62 medicines and medical supplies across all the public health facilities from October 2019 to November 2023 (see Table S5 for a list of products).

Supply data. We collect expiry date[8] and available supply units of all medicines and medical supplies per quarter from mSupply (*52*), a pharmaceutical logistics and warehouse management system used in Sierra Leone (and over 40 other countries). Prior to each allocation quarter, local staff are required to conduct stock counts and record the information in mSupply. In addition, we use invoice records from the mSupply system to determine the stock received by each public health facility to evaluate compliance for district-level allocations. Table S6 provides summary statistics on the supply of each product.

Catchment population. Estimating granular population is particularly challenging in developing countries. To address this, we leverage multiple publicly available datasets to estimate each health facility's catchment population (i.e., the number of people each health facility is expected to serve): the WorldPop Global Project Population Data (*53*), the global friction surface dataset (`Oxford/MAP/friction_surface_2019`) from Google Earth (*54*), and satellite imagery (*55*) also accessed through Google Earth. We utilize these to create population estimates for our catalytic priors (see §1.5.2).

[8] In line with the NMSA's existing practice, we prioritize allocating near-expiry products to larger districts, where higher consumption minimizes waste by ensuring supplies are used up prior to expiration.

1.2 Data Processing

Training data for demand prediction model. The data for training our demand prediction model is derived from the historical consumption data extracted from DHIS2. We use this time series data both to construct the demand $\xi^*_{t,n}$ in the current period t for facility n that we are trying to predict, as well as the features $x_{t,n}$ for prediction. For example, we construct facility-specific features, including: consumption, product, facility ID, facility type, latitude and longitude of the facility's geo-location, district, average consumption of the product for the facility in the past $\{1, 2, 3, 4, 5, 6\}$ months respectively, standard deviation of the consumption in the past 3 and 6 months respectively, total sample size for the facility-product pair, year, month, average consumption of the product across facilities in the past $\{1, 2, 3, 4, 5, 6, 10\}$ months respectively. These features were determined based on domain knowledge and feature engineering.

This data can sometimes be unreliable due to random or inconsistent data entry at individual facilities, requiring careful preprocessing. First, we exclude any observations where the inflow and outflow are inconsistent:

$$\begin{aligned}&\text{Closing Balance}\\ &\neq \text{Opening Balance} + \text{Quantity Received} - \text{Quantity Dispensed} + \text{Adjustment/Loss}\end{aligned}$$

Second, we remove observations where all recorded quantities were zero, since this likely indicates the use of a default value. Third, we exclude extreme outliers—specifically the top and bottom 5% of values, to mitigate the impact of potential data entry errors. Table S7 summarizes the percentage of data that is excluded due to data quality issues by product.

One major challenge is that we only observe consumption, which does not equal demand when there is a stockout; this issue is called *demand censoring* (*56*). To ensure we are predicting actual demand, we use a standard strategy where we drop censored observations (*57*). In particular, when constructing $(x_{n,t}, \xi_{n,t})$ pairs for training, we only include observations where

no stockout occurred; then, consumption is equal to the demand, so we can take it to be $\xi_{n,t}$. One limitation of this strategy is that it introduces *covariate shift*, since there may be systematic differences between time periods/facilities where stockouts occur and those where they do not occur. Covariate shift has the potential to degrade performance of the model compared to what is expected based on test set evaluation. Thus, we use unbiased population-based models that do not suffer from censoring as a catalytic prior when training our predictive models to mitigate the bias (see details in §1.5.2). We also perform several robustness checks to ensure the quality of our model, which are described in §2.4. Importantly, our econometric evaluation is performed using consumption outcomes alone, which are fully observed, so they are not affected by demand censoring (unlike our predictive model).

Panel data for evaluation. Our evaluation uses the same historical DHIS2 data as above but focuses on consumption, which is fully observable. As noted in the main paper, given a fixed budget, increasing consumption is mathematically equivalent to reducing unemet demand. We use data from 2022 Q3 to 2023 Q3. We start at 2022 Q3 since this is when the NMSA started using a standardized Excel allocation tool; prior to this period, allocation procedures were inconsistent, not allowing for reliable counterfactual estimation. We constructed a balanced panel dataset with 1,058 facilities for evaluation.

The scale of quarterly consumption at each facility depends on the type and size of the facility's catchment population. To account for these differences, we normalize each product's consumption at the facility level by subtracting the mean consumption of that product across all facilities and dividing by the standard deviation:

$$\text{NormalizedConsumption}_{n,m} = \frac{\text{Consumption}_{n,m} - \text{MeanConsumption}_{m}}{\text{StdConsumption}_{m}},$$

where n represents the facility, m represents the product, $\text{MeanConsumption}_{m}$ is the average consumption of product m across facilities, and $\text{StdConsumption}_{m}$ is the standard deviation

of product m across facilities. For each facility at each quarter, we then calculate the average normalized consumption across the products available at that facility as:

$$\text{FacilityAverageNormalizedConsumption}_n = \frac{\sum_{m \in \text{AvailableProducts}_n} \text{NormalizedConsumption}_{n,m}}{\text{AvailableProducts}_n}.$$

This approach ensures that the facility-level average consumption reflects the relative performance across products while accounting for the variability in the consumption level of different products.

1.3 Existing Allocation Approach

Each quarter, the National Medical Supplies Agency (NMSA) of Sierra Leone allocates approximately 70-100 free healthcare products specifically for women and children under five years old, with supply primarily dependent on international donations. Distribution happens quarterly, and is based on a centralized two-stage push system (*37*), where supplies move from the central government to districts, and then to local health facilities. We focus on a subset of 45 products that are regularly distributed—chosen in collaboration with NMSA officials prior to our deployment—ensuring sufficient historical data for model training.

Until the deployment of our tool in 2023 Q2, the process of computing the allocation of stock to each health facility relied primarily on a complex Excel tool. An important aspect of this computation was organizing health facilities into three administrative categories: District Medical Stores (DMS), District Hospitals (DH), and Western Area Hospitals (WAH). The DMS includes four facility types: Community Health Centers (CHC), Community Health Posts (CHP), Maternal and Child Health Posts (MCHP), and Clinics. Prior to each allocation cycle, all public health facilities submit requests based on their recent three-month rolling average of consumption. Healthcare workers can provide this information based on DHIS2, mSupply, or their own professional judgment.

Upon receiving all facility requests, the NMSA implements a structured allocation process. First, the NMSA determines the distribution proportions among the three primary healthcare

facility categories, typically 70% of total stock to DMS across 16 districts, 15% to DH, and 15% to WAH. Following this initial distribution, each district receives a specific allocation based on multiple criteria, including district population, poverty levels, product types, and submitted requests. For example, if a DMS in a particular district is allocated 10% of the DMS share, it receives a quantity calculated as

$$\text{total stock} \times 0.7 \times 0.1.$$

In cases where stock remains after the initial distribution and requests remain partially fulfilled, the NMSA makes further allocations based on the unfulfilled requests.

Once the proportion of allocations have been finalized, the distribution process follows a two-tier delivery system. The NMSA executes the "first mile" delivery to all districts, after which each district manages the "last mile" distribution to individual health facilities under the NMSA's guidance and supervision.

1.4 Supply Chain Optimization Algorithm

The key problem here is to allocate each limited medical resource to health facilities to minimize total unmet demand, defined as the number of patients turned away at facilities when supply was unavailable. This is a critical goal proposed by policymakers, as unmet demand reflects the quality of care provision and directly affects patient outcomes, particularly in regions with limited access to alternative care options.

We optimize the allocation of each product separately.[9] There is a fixed total budget $b \in \mathbb{R}$ to be distributed across $N \in \mathbb{N}$ facilities. Each facility $n \in [N]$ has an estimated demand Ξ_n, where $\Xi \in \mathbb{R}^N$ is a real-valued random vector with an estimated distribution $\mathbb{P}_\Xi$. We denote the allocation decision and stock on hand as $a \in \mathbb{R}^N$ and $s \in \mathbb{R}^N$ respectively, where a_n is

[9]Products are allocated independently since delivery trucks visit districts in a fixed sequence, preventing any inter-dependencies between product allocations due to truck capacity constraints.

the allocation intended for facility n; s_n is the stock on hand at each facility n. We can write the expected unmet demand, which measures the amount of unmet demand on average across facilities and over Ξ:

$$\mathbb{E}_{\Xi}\left[\sum_{n\in[N]}\max\{\Xi_n - a_n - s_n, 0\}\right]. \tag{S1}$$

If the total budget b is very large, we can choose all a_n sufficiently high to ensure that our objective is minimized at ≈ 0; if the total budget b is very low compared to the total demand $\sum_{n\in[N]}\Xi_n$, then most facilities will suffer stockouts, and we cannot do much better than $(\sum_{n\in[N]}\Xi_n) - b$. However, we find that the total budget is often on the order of the total demand (i.e., $b \approx \sum_{n\in[N]}\Xi_n$), likely because the budget is adjusted over time to meet demand. In this case, many facilities are over- and under-stocked; then, minimizing this objective requires setting the allocations a_n to be as close to the excess demand $\Xi_n - s_n$ as possible.

Note that we focus on allocating over a single period instead of multi-period allocation. This is because we find that forecasting demand beyond one quarter is far too noisy to be of value.

Optimization strategy. When Ξ is constant, the optimal policy can be straightforwardly expressed as a linear program. To account for the uncertainty in Ξ, we use *sample average approximation* (SAA), which takes K demand samples from the estimated demand distribution $\xi^{(k)} \sim \mathbb{P}_{\Xi}$ (for $k \in [K]$), and then optimizes the objective on average across these samples. The resulting optimization problem is

$$a^* = \underset{a\in\mathbb{R}^N}{\arg\min}\ \frac{1}{K}\sum_{k=1}^{K}\sum_{n=1}^{N} c_n^{(k)} \quad \text{subj. to} \quad c^{(k)} \geq \xi^{(k)} - a, \quad c \geq 0, \quad \sum_{n=1}^{N} a_n \leq b, \tag{S2}$$

where vector inequalities are element-wise, $c_n^{(k)}$ denotes the unmet demand for facility n in sample k, and b is the total available stock to be allocated. The first two constraints ensure $c_n^{(k)} = \max\{\xi_n^{(k)} - a_n, 0\}$, with one of these constraints necessarily binding to minimize the objective. The last constraint ensures the total allocation does not exceed the available stock.

Warehouse matching. Based on the final allocation generated by our model, we determine the specific inventory in warehouses that should be shipped to each facility. In line with their existing policy, we preferentially allocate faster-expiring stock to higher-volume facilities where it is more likely to be allocated prior to expiration. In particular, for each product, we first rank the stock based on time to expiration. Then, we iterate through facilities based on a ranking provided by the NMSA (typically, facilities with larger catchment populations are ranked higher). For each facility, we allocate stock that has the earliest expiry date across all warehouses, continuing until the facility's allocation is fully met.

1.5 Machine Learning Framework

In this section, we present our machine learning framework for demand predictions. We train a random forest, improving performance using standard methods from the literature on multi-task learning and catalytic priors; we then use a novel decision-aware learning approach to better align our predictions with the downstream optimization loss.

1.5.1 Multi-Task Learning

A common approach for demand estimation in supply chain management and global health is to fit a distribution to historical consumption patterns, e.g., one could estimate the mean and variance of each facility-product pair separately based on its historical data. Mathematically, it can be viewed as solving the following maximum likelihood problem:

$$\tilde{\ell}(\mu, \sigma) = -\sum_{t=1}^{T} \sum_{n=1}^{N} \log \mathcal{N}(\xi_{t,n}^*; \mu_n, \sigma_n^2),$$

where $\mathcal{N}(x; \mu_n, \sigma_n^2)$ is the Gaussian probability density function in x with mean μ_n and standard deviation σ_n and $\tilde{\ell}(\mu, \sigma)$ is the negative log-likelihood. This objective $\tilde{\ell}$ decomposes across facilities $n \in [N]$, and the solution for a given n is the empirical mean and variance. However, this strategy cannot learn dynamic patterns such as seasonal effects or demand that is elevated for

a period of time (e.g., due to an outbreak). Time series models like ARIMA are also infeasible due to the limited number of observations (on average, 28) we have for each facility-product pair—rather, we must leverage cross-facility and cross-product correlations.

Multi-task learning allows us to train a single model on multiple interrelated tasks (i.e., the different facility-product pairs). By aggregating data and transferring knowledge across related tasks, multi-task learning increases the effective sample size for each task (*26*). To this end, we first categorize all products into two types—medicines or medical supplies/equipment; we learn separate predictive models for these two categories. We associate each demand observation $\xi^*_{t,n}$ with a covariate vector $x_{t,n} \in \mathbb{R}^d$, which consists of features constructed from (i) the facility n, (ii) the time step t, and (iii) the historical demand over the k previous steps $\xi^*_{t-k,n}, \xi^*_{t-k+1,n}, \ldots, \xi^*_{t-1,n}$ (it also includes features of the product being allocated), based on domain knowledge and extensive feature engineering. Then, for each category (medicines or medical supply/equipment), we train a single model on the resulting dataset $\{(x_{t,n}, \xi^*_{t,n}\}_{t\in[T],n\in[N]}$. In particular, we seek to train predictors $\mu_\theta(x)$ and $\sigma_\theta(x)$ for the mean and standard deviation, respectively, where $\theta \in \Theta$ are the parameters, using the following objective:

$$\tilde{\ell}(\theta) = -\sum_{t=1}^{T}\sum_{n=1}^{N} \log \mathcal{N}(\xi^*_{t,n}; \mu_\theta(x_{t,n}), \sigma_\theta(x_{t,n})^2). \tag{S3}$$

In practice, we find that the following strategy works well. First, we use a random forest to fit the mean $\mu_\theta(x_{t,n})$, assuming the variance is constant. Then, we fit $\sigma_\theta(x_{t,n})$ based only on the historical data for the facility n and the current product in question.

To evaluate our multi-task learning strategy, we compare to two techniques that do not use multi-task learning. First, we consider the 3-month rolling average, which is the prediction strategy used by the existing Excel tool and more broadly is a common strategy for demand forecasting in LMICs (*21–23*). Second, we compare to a standard distribution modeling approach, where we fit a demand distribution for each facility-product pair based only on historical data

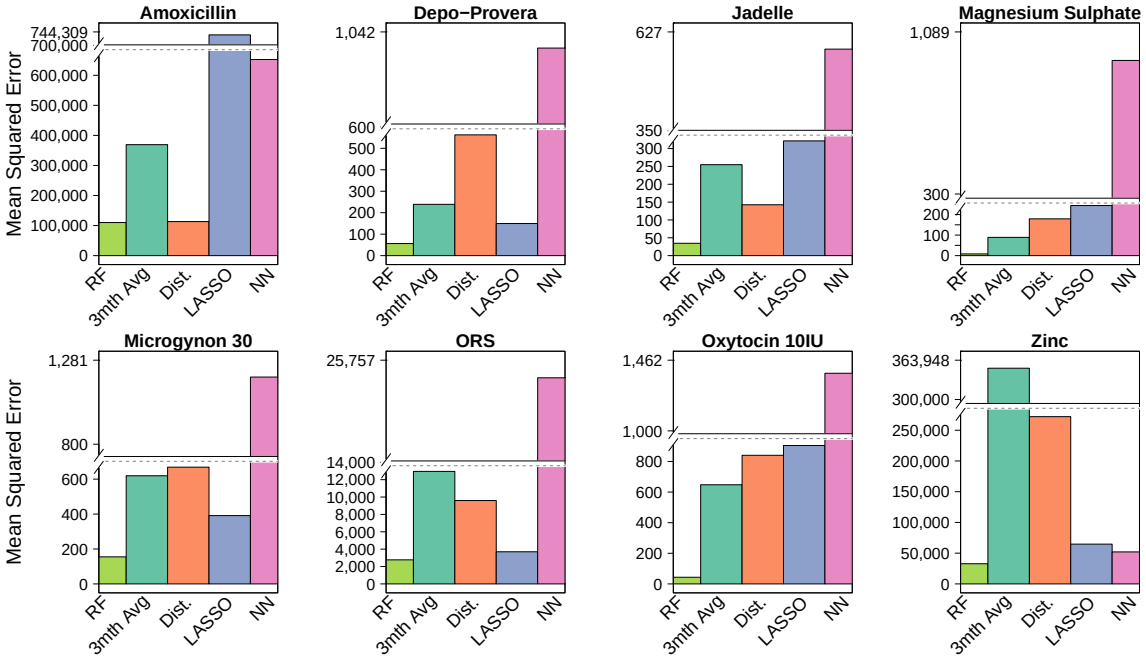

Figure S1: Prediction error comparison between different model families. Random forests (RF) yield the smallest prediction error for focal essential medicines chosen by the NMSA, compared to using a rolling 3-month average (3mth Avg), distribution modeling (Dist.) described in §1.5.1, LASSO regression (LASSO), and neural networks (NN).

from that facility-product pair. We examined the fit of 46 well-known candidate distributions (e.g., normal, beta, gamma, exponential) and chose the Nakagami distribution as the one that best minimized the sum of squared errors between the fitted probability density function (PDF) and a histogram of the historical data. We also compare to two other standard model families, LASSO regression and neural networks (NN), trained using the same multi-task learning pipeline as our random forest. In this comparison, we focus on evaluating the mean squared error (MSE) of μ_θ, since we fit σ_θ separately. Results are shown in Fig. S1; as can be seen, random forests consistently achieve lower MSE across all products on a held-out test set.

1.5.2 Catalytic Priors

We implement *catalytic priors* (*33*) as a safeguard to improve our model's robustness to inequities in data quality (arising from missing data or censoring). Oftentimes, low quality data come from poorer districts—thus, a model trained only on the available data may be biased in poorer districts, which may create unintentional disparities in predictive performance and downstream resource allocation. A natural strategy for mitigating such bias is to incorporate auxiliary data sources that are less likely to suffer from bias. For example, in public health, a standard approach is population-based resource allocation (PBRA), which uses population estimates to guide proportional resource allocation needs (*58*). Although population-based prediction is noisier (i.e., higher variance) because it cannot capture time-dependent patterns (e.g., seasonality of demand), it is less biased since the observations do not suffer from nonrandom missingness or censoring. Catalytic priors (*33*) allow us to suitably trade off bias and variance by regularizing our random forest with the simpler but less biased population-based model, which predicts demand only based on an estimate of the at-risk population in the catchment of a health facility.

For a given product, denote the prediction of the population-based model by $\xi^0_{n,t} = \mu_C(p_n) = r \cdot C \cdot p_n$, where p_n is our estimated population in the catchment of facility n; $r \in \mathbb{R}$ is an estimated multiplier derived from census data to account for the at-risk population (defined by the percentage of women and children in a given area); and $C \in \mathbb{R}$ is a single *product-specific* parameter estimated from our historical data (i.e., the average quarterly demand per unit of at-risk population). The catchment population p_n served by each health facility is not readily available, so we estimate it as follows:

1. First, we collect data on geographic coordinates of health facilities in Sierra Leone from several sources, including Google Maps and Geo-Referenced Infrastructure and Demographic Data for Development (GRID3) (*50*).

2. Next, we use Google Earth engine's satellite imagery datasets to compute the normalized difference vegetation index (NDVI) on a 10km×10km patch around each facility at monthly resolution between January 2022 and December 2022. NDVI serves as a proxy for vegetation density, which can indicate human activity.[10]

3. Then we use "friction surface" data from Google Earth (*54*) to obtain the travel time between every facility and pixel of the area with potential human activity. This allows us to define the catchment area based on minimal travel time.

4. Finally, we estimate p_n for health facility n by using data from WorldPop (*53*), which provides population count estimates for each 100m×100m grid cell.

We estimate r, the proportion of women and children, using 2015 Sierra Leone Census data (*59*) at the chiefdom-level.[11]

We then follow (*33*), generating synthetic observations from μ_C to act as a Bayesian prior for our random forest. In particular, we use μ_C to construct a single synthetic example $(x_{n,t}, \mu_C(p_n))$ for each facility $n \in [N]$ and time period $t \in [T]$. We then train $\mu_{\hat{\theta}}$ on a weighted combination of the original dataset and this synthetic dataset. Intuitively, the synthetic dataset naturally regularizes $\mu_{\hat{\theta}}$ towards a stable estimate μ_C in data-poor regions of the covariate space.

1.5.3 Decision-Aware Learning

The next step is to incorporate our predictions into the optimization model to generate allocation decisions. However, the model is usually trained to forecast demand using a standard objective such as mean-squared error (MSE), which focuses on minimizing prediction error and ignores

[10]NDVI is derived from satellite images using the formula: NDVI $= (\text{NIR} - \text{red})/(\text{NIR} + \text{red})$. Since vegetation reflects light in the near-infrared (NIR) spectrum and absorbs light in the red spectrum, areas with higher photosynthesis activity exhibit larger NDVI values.

[11]Chiefdoms are a more granular administrative unit than districts; there are 190 chiefdoms in Sierra Leone. We further verified these estimates using data from the United Nations Office for the Coordination of Humanitarian Affairs OCHA (*60*)

the decision error in the downstream optimization problem, making it *decision-blind*. This can result in poor performance since it may not focus the capacity of the machine learning model on predictions that are actually relevant to making decisions (*28, 29*). Thus, there has been interest in algorithms that incorporate the *decision loss* into the training algorithm (*29, 61*). We found that existing decision-aware learning algorithms were either computationally intractable at our scale or incompatible with the rest of our prediction and optimization pipeline; thus, we develop a novel and light-weight decision-aware learning approach—it relies only on re-weighting observations in model training, which can be easily integrated with existing data pipelines.

In our setting, the decision loss is

$$\ell(\hat{\mu}; \mu) = L(a^*(\hat{\mu}), \mu^*),$$

where

$$L(a; \mu) = \sum_{n=1}^{N} \mathbb{E}_{\Xi_n \sim \mathcal{N}(\mu_n, \sigma^2)} \left[\max\{\Xi_n - a_n - s_n, 0\} \right],$$

is the unmet demand for allocation $a \in \mathbb{R}^N$ assuming the true demand for facility $n \in [N]$ is $\mathcal{N}(\mu_n, \sigma^2)$ (recall that for training the random forest, we have assumed that the standard deviation is a fixed value σ), and where

$$a^*(\mu) = \arg\min_{a \in \mathbb{R}^N} L(a; \mu)$$

is the optimal allocation assuming the demand distributions for $n \in [N]$ is $\mathcal{N}(\mu_n, \sigma^2)$. In other words, the decision loss is the expected unmet demand incurred when using the predictions $\hat{\mu}$ in our optimization problem. To address objective mismatch, we could train μ_θ to directly minimize the decision loss:

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \sum_{t=1}^{T} \sum_{n=1}^{N} \ell(\mu_\theta(x_{t,n}); \mu^*_{t,n}).$$

Algorithms for doing so have been proposed in the setting of linear regression (*28*), and in the more general setting of differentiable model families by taking gradients through the optimization problem (*30, 31*). However, existing techniques are often limited to specific prediction setups or become computationally intractable for large-scale problems (*30, 31, 43*).

Our strategy is to Taylor expand the optimal decision loss, which we will show can be interpreted as up-weighting data more relevant to the downstream optimization problem. This approach can also be easily integrated with existing pipelines and is flexible enough to handle a broad range of model families. In particular, we approximate the decision loss by Taylor expanding it around $\hat{\mu} - \mu_0$ (where μ_0 is the current prediction and $\hat{\mu} = f_\theta(x)$), yielding

$$L(a^*(\hat{\mu}); \mu^*) \approx L(a^*(\mu_0); \mu^*) + \nabla_a \ell(a^*(\mu_0); \mu^*)^\top \nabla_\mu a^*(\mu_0)^\top (\hat{\mu} - \mu_0). \tag{S4}$$

Since the first term is a constant, we can ignore it; in particular, we have

$$\begin{aligned}
\ell(\hat{\mu}; \mu^*) &\approx \nabla_a L(a^*(\mu_0); \mu^*)^\top \nabla_\mu a^*(\mu_0)^\top (\hat{\mu} - \mu_0) + \text{const} \\
&= \nabla_a L(a^*(\mu_0); \mu^*)^\top \nabla_\mu a^*(\mu_0)^\top (\hat{\mu} - \mu^*) + \text{const}.
\end{aligned}$$

Here, we have replaced μ_0 with μ^*; since both of these are constants, it does not affect the optimal solution. With this replacement, we can upper bound the term by its absolute value to avoid "overshooting" ξ^*, resulting in a weighted absolute error loss:

$$\begin{aligned}
\ell(\hat{\mu}; \mu^*) &= \sum_{t=1}^{T} \sum_{n=1}^{N} w_{t,n} (\hat{\mu}_{t,n} - \mu^*_{t,n}) + \text{const} \\
&\leq \sum_{t=1}^{T} \sum_{n=1}^{N} |w_{t,n}| \cdot |\hat{\mu}_{t,n} - \mu^*_{t,n}| + \text{const},
\end{aligned}$$

where

$$w_{t,n} = \left(\nabla_\mu a^*(\mu_0)^\top \nabla_a L(a^*(\mu_0); \mu^*) \right)_{t,n}$$

Our algorithm uses this upper bound as the loss function, which works with any standard machine learning algorithm that can take weighted examples. In particular, we train our random forest to minimize this loss on the training data:

$$\hat{\theta} = \arg\min_{\theta} \sum_{t=1}^{T} \sum_{n=1}^{N} |w_{t,n}| \cdot |\mu_\theta(x_{t,n}) - \xi^*_{t,n}|.$$

Note that we do not observe the true demand $\mu^*_{t,n}$, so we use $\xi^*_{t,n}$ as an estimate. Finally, as a heuristic, we replace the absolute error with the squared error, which is more computationally tractable.

An important insight here is that the weight can be interpreted as re-weighting training examples. The w_n can be decomposed as two gradients, and this can be computed numerically efficiently for a general class of convex programs (*62*). In our case, we can derive the weights analytically by solving the optimization problem:

$$\begin{aligned} a^*(\mu) = &\arg\min_{a \in \mathbb{R}^N} \mathbb{E}_{\Xi_n \sim \mathcal{N}(\mu_n, \sigma^2)} \left[\max\{\Xi_n - a_n - s_n, 0\}\right] \\ &\text{subj. to} \quad a_n \geq 0 \ (\forall n \in [N]), \quad \sum_{n=1}^{N} a_n \leq b, \end{aligned}$$

Letting $\Xi_n = \mu_n + \eta_n$, where $\eta_n \sim \mathcal{N}(0, \sigma^2)$ i.i.d., we can form the Lagrangian:

$$L(a, \lambda) = \sum_{n=1}^{N} \mathbb{E}_{\eta_n} \left[\max\{\mu_n + \eta_n - a_n - s_n, 0\}\right] + \lambda_0 \left(b - \sum_{n=1}^{N} a_n\right) + \sum_{n=1}^{N} \lambda_n a_n.$$

The KKT conditions are

$$\begin{aligned} 0 &= \nabla_{a_n} L(a^*(\mu), \lambda^*(\mu)) = -\mathbb{P}_{\eta_n} \left[a^*_n(\mu) \leq \mu_n + \eta_n - s_n\right] + \lambda^*_0(\mu) + \lambda^*_n(\mu) \\ 0 &= \nabla_{\lambda_0} L(a^*(\mu), \lambda^*(\mu)) = \sum_{n=1}^{N} a^*_n(\mu) - b \\ 0 &= \lambda^*_n(\mu) a^*_n(\mu) \\ 0 &\leq \lambda^*_n(\mu) \\ 0 &\leq a^*_n(\mu) \end{aligned}$$

Let

$$\mathcal{I}(\mu) = \{n \in [N] \mid \lambda_n^*(\mu) = 0\}$$

Note that if $n \notin \mathcal{I}(\mu)$, then $a_n^*(\mu) = 0$, so the first condition becomes

$$a_n^*(\mu) = \begin{cases} \mu_n - s_n + F_{\eta_n}^{-1}(1 - \lambda_0^*(\mu)) & \text{if } n \in \mathcal{I}(\mu) \\ 0 & \text{otherwise,} \end{cases}$$

where F_{η_n} is the CDF of η_n. Since the η_n are i.i.d., we write this CDF as simply F_η. Summing over $n \in [N]$, we have

$$\begin{aligned} b = \sum_{n=1}^{N} a_n^*(\mu) &= \sum_{n \in \mathcal{I}(\mu)} \left(\mu_n - s_n + F_\eta^{-1}(1 - \lambda_0^*(\mu))\right) \\ &= \left(\sum_{n \in \mathcal{I}(\mu)} \mu_n - s_n\right) + |\mathcal{I}(\mu)| \cdot F_\eta^{-1}(1 - \lambda_0^*(\mu)). \end{aligned}$$

Thus, we have

$$F_\eta^{-1}(1 - \lambda_0^*(\mu)) = \frac{b - \sum_{n \in \mathcal{I}(\mu)} (\mu_n - s_n)}{|\mathcal{I}(\mu)|},$$

so

$$a_n^*(\mu) = \begin{cases} \mu_n - s_n + \frac{1}{|\mathcal{I}(\mu)|} \left(b - \sum_{m \in \mathcal{I}(\mu)} (\mu_m - s_m)\right) & \text{if } n \in \mathcal{I}(\mu) \\ 0 & \text{otherwise,} \end{cases}$$

Taking the derivative with respect to μ, we have

$$\nabla_{\mu_m} a_n^*(\mu) = \delta_{m,n} - \frac{1}{|\mathcal{I}(\mu)|}$$

Thus, we have

$$\begin{aligned}
\nabla_{\mu_m} L(a^*(\mu), \lambda^*(\mu)) &= \nabla_{\mu_m} \sum_{n=1}^{N} \mathbb{E}_{\eta_n} \left[\max\{\mu_n + \eta_n - a_n^*(\mu) - s_n, 0\}\right] \\
&= \sum_{n=1}^{N} \mathbb{P}_{\eta_n} \left[a_n^*(\mu) \leq \mu_n + \eta_n - s_n\right] \nabla_{\mu_m} a_n^*(\mu) \\
&= \sum_{n=1}^{N} \mathbb{P}_{\eta_n} \left[a_n^*(\mu) \leq \mu_n + \eta_n - s_n\right] \left(\delta_{m,n} - \frac{1}{|\mathcal{I}(\mu)|}\right) \\
&= \mathbb{P}_{\eta_m} \left[a_m^*(\mu) \leq \mu_m + \eta_m - s_m\right] + \frac{1}{|\mathcal{I}(\mu)|} \sum_{n=1}^{N} \mathbb{P}_{\eta_n}[a_n^*(\mu) \leq \mu_n + \eta_n - s_n] \\
&= \mathbb{P}_{\eta_m} \left[a_m^*(\mu) \leq \mu_m + \eta_m - s_m\right] + \text{const} \\
&\approx \mathbb{I}[a_m^*(\mu) \leq \mu_m - s_m] + \text{const},
\end{aligned}$$

where the approximation on the last line holds when σ is small. Using this gradient, our predictive model's objective can be approximated as

$$\hat{\theta} = \arg\min_{\theta} \sum_{t=1}^{T} \sum_{n=1}^{N} (\mathbb{I}[a_{t,n}^*(\mu) \leq \mu_{t,n} - s_{t,n}] + c) \cdot |\mu_\theta(x_{t,n}) - \xi_{t,n}^*|$$

for some constant c. This loss up-weights the training examples that are more likely to experience unmet demand. One remaining issue is that we do not know $a_{t,n}^*(\mu)$, which is required to compute the weight. We approximate it by first training a decision-blind model, using it to optimize the allocations, and then using these allocations to determine the weights.

Now, we compare the out-of-sample performance of our decision-aware learning approach against the following baselines:

- **Excel tool:** Tool previously used by the NMSA to make allocations.
- **Decision-blind ablation:** Follows our pipeline but uses the MSE loss to train the random forest instead of our decision-aware learning algorithm.
- **StochOptForest:** Follows our pipeline but trains decision-aware random forests using the end-to-end optimization strategy by (*43*).

- **Distribution modeling:** Estimates demand via distribution modeling (as described in §1.5.1), and then optimizes allocations based on these forecasts.
- **Global Health:** Estimates demand via a 3-month rolling average (described in §1.5.1, common in global health (*21–23*)), and then optimizes allocations based on these forecasts.
- **Population-based:** Allocate total stock to chiefdoms proportionally to their at-risk population (women and children), as commonly done in global health (*22*). Within a chiefdom, all facilities are treated equally.[12]

We apply our framework and evaluate the unmet demand for each facility-product pair:

$$\text{UnmetDemand}_{i,j} = \begin{cases} 0 & \text{if } \mu_{i,n} - a_{i,n} - s_{i,n} < 0 \\ \mu_{i,n} - a_{i,n} - s_{i,n} & \text{otherwise} \end{cases} \quad (\forall \text{products } i,\ \text{facilities } n) \tag{S5}$$

where $a_{i,n}$ is allocation decision and $\mu_{i,n}$ is the true demand. Then, we compute the total unmet demand across all facilities and divide it by the total consumption to obtain a normalized total unmet demand, which we average across products:

$$\text{NormalizedTotalUnmetDemand} = \frac{1}{\#\text{ products}} \sum_i \frac{\sum_n \text{UnmetDemand}_{i,n}}{\sum_n \mu_{i,n}}. \tag{S6}$$

To test allocation performance in challenging environments, we focus our evaluation on lower budgets—specifically, we use the 25th percentile of quarterly budgets (by product) observed in our data. We then measure how much our approach reduces unmet demand compared to the baseline:

$$\frac{\text{NormalizedTotalUnmetDemand}_{\text{Baseline}} - \text{NormalizedTotalUnmetDemand}_{\text{Ours}}}{\text{NormalizedTotalUnmetDemand}_{\text{Baseline}}}. \tag{S7}$$

Our results are in Table S1, illustrating that our approach outperforms other baselines. It performs significantly better than non-machine learning approaches, demonstrating the predictive

[12]Note that current data sources do not provide facility-level population estimates, which we construct from a variety of data sources in §1.5.2.

Table S1: **Average % Improvement in Unmet Demand of Our Framework vs. Baselines**

Method	Improvement %
Our Framework	0%
Decision-Blind Ablation	5%
Population Based Census	27%
Distribution Modeling	82%
Global Health (3 Month Rolling Avg)	88%
StochOptForest	92%
Existing Excel Tool	98%

power of machine learning. Next, StochOptForest most likely performs poorly since it is unable to integrate its decision-aware learning strategy with multi-task learning. At a high level, their algorithm assumes that there is a single optimization problem for each example in the dataset (i.e., one for each facility-product pair). However, in our problem, a single product's optimization problem is associated with many examples (i.e., observations across all facilities). Thus, to apply their approach, we need to decouple the optimization problem into a separate optimization problem for each facility, which we do using the optimal dual variable λ. However, this eliminates cross-learning between facilities, which may explain the poor results. Finally, we also demonstrate a 5% improvement compared to a purely decision-blind approach; while this improvement is comparatively smaller, a 5% reduction in unmet demand still has significant implications for social welfare. Furthermore, it demonstrates the potential for decision-aware learning to improve performance even compared to powerful state-of-the-art models such as multi-task random forests.

2 Evaluation and Deployment

We provide details on our deployment (§2.1) and our main analysis using SynthDiD (§2.2). Then, we investigate real-world compliance to our allocations and show how it affects outcomes (§2.3).

Finally, we conduct multiple robustness checks to validate our main findings (§2.4).

2.1 Deployment Details

As noted in the main paper, the Sierra Leone national government deployed our system in 2023 Q3 for five randomly selected districts: Tonkolili, Falaba, Karene, Kono, and Pujehun. Prior to the deployment, the government had established predetermined supply levels for this period and allocated resources to control districts. The remaining supply was then assigned to the treatment districts, maintaining the independence of supply quantities between the two groups.

Before the implementation, we first conducted two training sessions for policymakers and frontline workers to provide them with a technical understanding of how our tool operated and what it did. We ensured that our allocation tool was compatible with the same formatted inputs and outputs as the prior Excel allocation tool, allowing users to maintain their existing workflows with minimal changes. A screenshot of our web interface is shown in Fig. S2.

The rollout of the deployment began in June 2023. The implementation timeline proceeded as follows: in the Falaba and Tonkolili districts, last-mile delivery to local health facilities was completed by mid-July, while in the Karene, Kono, and Pujehun districts, implementation extended to the end of July due to logistical delays caused by the presidential election in Sierra Leone. To evaluate the impact of the intervention, we primarily analyzed outcomes from 2023 Q2 to Q3. The government did not conduct an allocation in Q4 due to insufficient supply.

2.2 Main Analysis

We describe how we use SynthDiD (*45*) to analyze the impact of our deployment on patient consumption. SynthDiD identifies causal effects by ensuring that the difference between treated units and synthetic control units remain stable before treatment. To achieve this, SynthDiD assigns two sets of weights: one set to control units, and another set to time periods. The

Figure S2: **Our System's Web App Interface**: Consistent with their prior workflow, users upload an Excel sheet with stock information (which is directly downloaded from the mSupply database) and then run the tool to obtain downloadable allocation results. This web app is now owned and operated by the Sierra Leone national government.

control unit weights align the weighted average of control units' outcomes with the unweighted average of treated units' outcomes in the pre-treatment period, while the time period weights ensure that the weighted outcomes of these control units in the pre-treatment period closely match their unweighted outcomes in the post-treatment period. By combining these weights, SynthDiD constructs a synthetic comparison group whose pre-treatment trends align with those of the treated units, thereby providing a credible estimate of the treatment's causal impact. SynthDiD remains robust even when treatment and control groups show different trends before the intervention (unlike DiD, which requires parallel trends), and it can effectively control for

variations in outcomes that arise from both time-related and unit-specific factors (unlike synthetic controls).

Using unit weights $\hat{\omega}_i^{\text{sdid}}$ and time weights $\hat{\lambda}_t^{\text{sdid}}$ derived from Equations (4) and (6) in (*45*), the average effect of treatment on the treated (ATT) is estimated as follows:

$$\left(\hat{\tau}^{\text{sdid}}, \hat{\mu}, \hat{\alpha}, \hat{\beta}\right) = \underset{\tau,\mu,\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T} \left(Y_{it} - \mu - \alpha_i - \beta_t - W_{it}\tau\right)^2 \hat{\omega}_i^{\text{sdid}} \hat{\lambda}_t^{\text{sdid}} \tag{S8}$$

where Y_{it} represents the observed outcome for unit i at time t, with a baseline mean outcome μ, unit-specific fixed effects α_i, and time-specific fixed effects β_t. The treatment assignment indicator W_{it} determines whether unit i receives treatment at time t, with the corresponding treatment effect denoted by τ. To account for the differences between treated and control units over time, control unit weights $\hat{\omega}_i^{\text{sdid}}$ and time period weights $\hat{\lambda}_t^{\text{sdid}}$ are applied, ensuring an appropriate balance in estimating the treatment effect.

We independently perform our analysis using both the *synthdid* package in R and the *sdid* package in STATA, finding consistent results. We estimate standard errors using the jackknife (Algorithm 3 in (*45*)). To validate our use of SynthDiD, we also perform a standard event study, which shows that there are no statistically significant differences between treated and control units prior to our intervention and that the change in consumption emerges only after our system was deployed (see Fig. S3). We use Equation 8 in (*48*), which compares the treated-minus-synthetic-control difference in each time period t to a baseline pre-treatment difference:

$$(\bar{Y}_t^{Tr} - \bar{Y}_t^{Co}) - (\bar{Y}_{\text{baseline}}^{Tr} - \bar{Y}_{\text{baseline}}^{Co}),$$

where $\bar{Y}_{\text{baseline}}^{Tr}$ and $\bar{Y}_{\text{baseline}}^{Co}$ are the baseline means for the treated group and the synthetic control group, respectively. Unlike conventional event studies that choose a single pre-treatment period as the baseline, the SynthDiD framework selects the optimal pre-treatment weights $\hat{\lambda}_t^{sdid}$

$$\bar{Y}_{\text{baseline}}^{Tr} = \sum_{t=1}^{T_{\text{pre}}} \hat{\lambda}_t^{sdid} \bar{Y}_t^{Tr} \quad \text{and} \quad \bar{Y}_{\text{baseline}}^{Co} = \sum_{t=1}^{T_{\text{pre}}} \hat{\lambda}_t^{sdid} \bar{Y}_t^{Co}.$$

Then, we construct the event-study estimates from SynthDiD using the following steps:

1. **Initial Estimation:**
   (a) Fit SynthDiD on the full sample to obtain time period weights $\hat{\lambda}$ and the pre-treatment difference in outcomes.
   (b) Adjust post-treatment differences by subtracting the pre-treatment mean difference.
2. **Bootstrap Inference:**
   (a) For $b \in \{1, \ldots, B\}$, resample the data and re-estimate SynthDiD.
   (b) Compute the bootstrap-adjusted difference series for each replication.
   (c) Use these bootstrap replicates to compute standard errors and form confidence intervals.

We find there are no statistically significant differences before treatment and that the change in outcome emerges only after treatment is introduced, validating our estimation strategy.

2.3 Compliance Analysis

Our system was implemented as a decision support tool that could be overridden by district pharmaceutical managers. Thus, we examine compliance with our recommendations in 2023 Q2, and its impact on consumption outcomes. This is notable because, after the government rolled out our system to the entire country in Q3, all districts reported full compliance.

Measuring compliance. To measure compliance, we manually collected the local pharmacist's allocation document and cross-checked all the invoices pulled from mSupply. To quantify compliance of a treated district, we normalize the allocation quantity for each facility-product pair in that district, calculate the absolute difference between the actual allocation and our

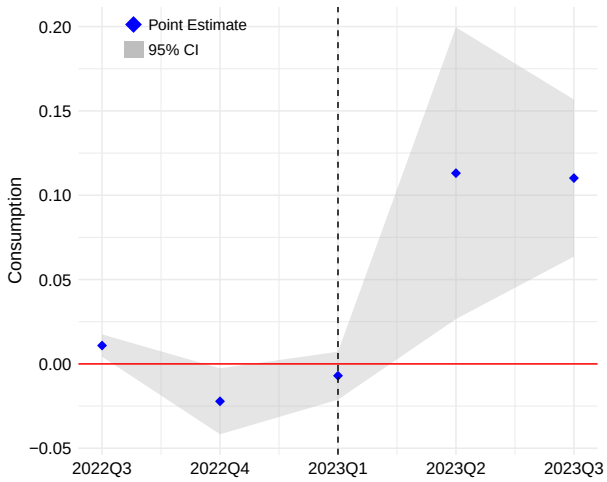

Figure S3: **Event Study of Q2 deployment.**This plot shows the estimated treatment effects across time. Blue diamonds represent point estimates, while the shaded region denotes the 95% confidence interval. The solid vertical line indicates the time of deployment. As expected, the pre-treatment estimates are close to zero, while the post-treatment estimates increase over time, indicating a positive impact on consumption attributable to our deployment.

Table S2: **Compliance of Treated Districts in 2023 Q2**

District	Normalized Avg Absolute Diff
Tonkolili	0.000
Falaba	0.028
Karene	0.039
Kono	0.073
Pujehun	0.109

deployed suggestion, and then average this value across facility-product pairs. As can be seen from Table S2, the Kono and Pujehun districts have significantly lower compliance than the other three districts. This is likely due to logistical and communication issues that arose during the implementation of our system in 2023 Q2, which were resolved soon after, yielding perfect compliance in Q3 and beyond.

Impact on consumption. According to the above analysis, we classify Tonkolili, Falaba, and Karene as compliers, and Kono and Pujehun as non-compliers. To estimate the Local Average Treatment Effect (LATE), we use an instrumental variable (IV) approach (*63*), with the government's random treatment assignment as the instrument. The treatment effect is estimated using Two-Stage Least Squares (2SLS), with the first-stage regression specified as

$$D_i = \pi_1 Z_i + \pi_2 X_i + \mu_i, \tag{S9}$$

where D_i is the compliance dummy, Z_i is the treatment assignment, and X_i denotes covariates (including indicators for the district, quarter, and facility type). This regression captures the proportion of compliers. The second-stage regression is given by

$$Y_i = \beta_1 D_i + \beta_2 X_i + \epsilon_i, \tag{S10}$$

where Y_i is the outcome variable (i.e., normalized consumption). Table S3 shows the results, which are consistent with our main analysis.

2.4 Robustness Checks

Next, we describe our robustness checks; our results are summarized in Table S4.

Difference-in-Differences (DiD). We use our panel data to conduct a DiD analysis (*64*), a standard causal inference method that estimates treatment effects by comparing changes in outcomes over time between the treated group and control group. This regression is specified as:

$$Y_{it} = \mu + \alpha_i + \beta_t + \tau(\text{Treat}_i \times \text{Post}_t) + \delta X_{it} + \epsilon_{it}. \tag{S11}$$

Y_{it} is the observed outcome (i.e., normalized consumption) for unit i at time t; μ denotes the mean outcome; α_i represents unit fixed effects; β_t represents time fixed effects; X_i denotes observed covariates (including indicators for the district and facility type); ϵ_{it} is the error term.

Table S3: **LATE IV Result**

	(1) OLS	(2) First-Stage	(3) IV
Dependent variable:	Consumption	Treated Complier	Consumption
Treated Complier	-0.045	–	**0.362*** **
	(0.011)		(0.044)
IV (treatment assignment)	–	0.447***	–
		(0.004)	
District fixed effect	YES	YES	YES
Quarter fixed effect	YES	YES	YES
Facility Type fixed effect	YES	YES	YES
Observations	217,330	217,330	217,330
R^2	0.003	–	0.099
$F - Stat$	–	14,074	–

**p<0.05, **p<0.01, ***p<0.001.*
Notes: Significance level stays and coefficient is larger for IV result without fixed effects.

Treat_i is a dummy indicator of treatment; Post_t is a dummy indicator for post-treatment periods; and, τ captures the desired treatment effect. We find that the magnitude and significance of the increase in consumption are similar to those obtained using SynthDiD.

Imputation strategies. In our main analysis, we excluded missing observations. We consider several alternative strategies based on imputing missing outcomes instead of dropping them:

- **Low-rank imputation (ImputedLowRank):** We use low-rank matrix completion (*49*), a standard approach for handling missing data in large matrices. We use the *softImpute* R package with rank $= 2$ and regularization parameter $\lambda = 0.1$ to impute missing consumption values.

- **Population-based imputation (ImputedPop):** We first estimate demand in proportion to each facility's catchment population, and then impute consumption as the minimum of the estimated demand and the computed allocation. (We compute the allocation via the Excel

tool in quarters prior to our tool's deployment, and using our tool otherwise).

- **Average imputation (ImputedAvgConsump).** Third, using comprehensive data on unique quarter, facility, and product pairs, we impute missing consumption values by assigning it to be the average quarterly consumption of the product across all facilities.

We find that our estimates are qualitatively similar to our main empirical evaluation, suggesting that our results are robust to different imputation strategies.

Alternative control group (Alt. Control). We also perform our analysis using an alternative control group of 25 products that were concurrently allocated using a different, pre-existing mechanism (see Table S5 for a list). The consumption levels for these products can be used as a control group throughout our study period across all districts in the country using a staggered treatment—i.e., an advantage of this analysis is that it can be performed not just for the partial deployment in 2023 Q2, but also for the nationwide implementation starting in Q3. We use the *sdid* package in STATA (*48*) to analyze our staggered implementation. We again find a statistically significant increase in consumption of 40% ($p < 0.01$).

Stockouts. We also assess the impact of our system on the number of facility-product stockouts. Note that reducing stockouts is not necessarily correlated with reducing unmet demand—for instance, we can reduce stockouts by allocating zero supply to a small number of high-volume facilities, which would ensure no stockouts at other facilities, but this approach would increase unmet demand. Our analysis shows a small decrease in stockouts ($p \approx 0.07$) but it is not statistically significant.

Table S4: **Robustness Results.** Column (1) shows DiD results; columns (2)-(4) show the SynthDiD results under different imputation strategies detailed in §2.4; columns (5)-(6) show SynthDiD results for stockout outcomes and under-served facilities, respectively; column (7) shows SynthDiD results using alternative controls in our staggered rollout.

	(1) DiD	(2) ImputedLowRank	(3) ImputedAvgConsump	(4) ImputedPop
Treated	0.124***	0.037***	0.068***	0.076***
	(0.038)	(0.013)	(0.031)	(0.028)
Observations	5,290	5,460	5,460	5,460

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Our balanced dataset includes 1,058 facilities across five quarters from 2022Q3.

	(5) Stockout	(6) Under-Served	(7) Alt. Control
Treated	-0.280	0.211***	0.383***
	(0.179)	(0.058)	(0.032)
Observations	5,290	4,055	54,465

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table S5: **Summary of Products Studied.** Treatment is an indicator of whether the product was allocated under our system. The mean, median, and standard deviation columns represent statistics for monthly facility-level consumption.

Product	Treatment	Mean	Median	Std Dev
Albendazole 400mg, Tab	1	299	200	402
Aluminium Hydroxide 500mg, Tab	1	260	100	377
Amoxicillin 250mg, Dispersible, Tab	0	1250	950	1474
Ampicillin 500mg, Pdr for IM/IV, Inj, Vial	1	137	66	441
Apron, Plastic, Disposable, Pcs	1	42	20	180
Bandage, Elastic, 8cm x 4m, Roll	0	14	2	48
Benzyl Benzoate 25%, Emulsion, 100ml, Bot	1	20	2	81
Blade, Surgical, No. 22, Sterile, Disposable, Pcs	0	34	10	122
Cannula, IV, 18G, Short, Sterile, Disposable, Pcs	1	27	15	81
Cannula, IV, 22G, Short, Sterile, Disposable, Pcs	0	23	10	80
Cannula, IV, 24G, Short, Sterile, Disposable, Pcs	1	33	20	122
Ceftriaxone 1g, Pdr for Inj, Vial	0	288	30	1022
Ceftriaxone 250mg, Pdr for Inj	1	233	98	355
Chlorhexidine Gluconate 5%, Solution, 1000ml, Bot	0	12	1	168
Chlorhexidine Gluconate 7.1%, Gel	1	7	4	25
Ciprofloxacin 500mg, Tab	1	285	130	429
Cloxacillin 500mg, Tab/Cap	0	416	100	825
Condom, Male	0	244	144	469
Copper-Containing Device (Copper T or Copper 7 or IUD)	0	15	1	44
Cotton Wool, Absorbent, 500g, Roll	0	5	2	27
Envelope, Dispensing, Plastic, 10cm x 7cm, Pcs	1	264	200	334
Epinephrine HCl (Adrenaline) 1mg/ml, Inj, 1ml, Amp	0	16	1	100
Erythromycin 125mg/5ml, Pdr for Susp, 100ml, Bot	1	204	50	428
Ferrous Sulphate 200mg, Tab	1	787	500	994
Folic Acid 5mg, Tab	1	1107	919	1389
Glove, Exam, Latex, Medium, Nonsterile, Disposable, Pcs	1	347	200	672
Glove, Surgical, Size 7.5, Sterile, Disposable, Pair	0	82	30	167
Glove, Surgical, Size 8, Sterile, Disposable, Pair	0	56	10	152
Glucose (Dextrose) 5%, IV Inj, 500ml, Soft Bag	1	71	14	223
Glucose (Dextrose) Hypertonic 50%, IV Inj, 50ml, Bot	1	19	4	51
IV Giving Set, Pcs	1	40	25	103
Jadelle	0	17	10	35
Levonorgestrel (Emergency Contraceptive) 1.5mg, Tab	0	10	3	25
Levoplant	0	11	4	28
Methyldopa 250mg, Tab	1	168	100	279
Metoclopramide HCl 10mg, Tab	1	305	100	545
Misoprostol 200mcg, Tab	1	65	20	258
Needle, Hypodermic, Luer, 21G, Sterile, Disposable	0	125	100	282
Needle, Hypodermic, Luer, 23G, Sterile, Disposable	0	119	100	198
Syringe, Luer, 10ml, Disposable, Pcs	1	67	30	380
Syringe, Luer, 20ml, Disposable, Pcs	1	47	20	131
Zinc Sulphate 20mg, Dispersible, Tab	1	527	350	793

Table S6: **Descriptive Statistics of the Supply.** We report supply quantities broken down between pre- and post- deployment of our system in 2023 Q2.

Product, unit	Average Pre-treated	Average Post-treated
Albendazole 400mg, Tab	2,486,700	1,106,200
Aluminium Hydroxide 500mg, Tab	–	73,500
Amoxicillin 250mg, Dispersible, Tab	1,872,250	2,069,600
Ampicillin 500mg, Pdr for IM/IV, Inj, Vial	738,850	896,975
Apron, Plastic, Disposable, Pcs	44,107	7,400
Bandage, Elastic, 8cm x 4m, Roll	–	1,400
Benzyl Benzoate 25%, Emulsion, 100ml, Bot	–	6,752
Blade, Surgical, No. 22, Sterile, Disposable, Pcs	400	2,250
Cannula, IV, 18G, Short, Sterile, Disposable, Pcs	194,450	47,100
Cannula, IV, 22G, Short, Sterile, Disposable, Pcs	–	312,100
Cannula, IV, 24G, Short, Sterile, Disposable, Pcs	341,175	289,000
Ceftriaxone 1g, Pdr for Inj, Vial	58,450	285,710
Ceftriaxone 250mg, Pdr for Inj	–	125,988
Chlorhexidine Gluconate 5%, Solution, 1000ml, Bot	14,378	2,612
Chlorhexidine Gluconate 7.1%, Gel	121,697	239,115
Ciprofloxacin 500mg, Tab	–	460,310
Condom, Male	6,703,200	11,789,280
Cotton Wool, Absorbent, 500g, Roll	–	4,527
Envelope, Dispensing, Plastic, 10cm x 7cm, Pcs	–	1,778,450
Epinephrine HCl (Adrenaline) 1mg/ml, Inj, 1ml, Amp	44,845	13,610
Erythromycin 125mg/5ml, Pdr for Susp, 100ml, Bot	–	146,093
Ferrous Sulphate 200mg, Tab	–	1,369,000
Folic Acid 5mg, Tab	7,493,500	2,203,000
Glove, Exam, Latex, Medium, Nonsterile, Disposable, Pcs	1,665,250	2,530,450
Glove, Surgical, Size 8, Sterile, Disposable, Pair	161,475	100,575
Glucose (Dextrose) 5%, IV Inj, 500ml, Soft Bag	35,667	200,660
Glucose (Dextrose) Hypertonic 50%, IV Inj, 50ml, Bot	165,000	85,737
IV Giving Set, Pcs	16,650	416,437
Levonorgestrel (Emergency Contraceptive) 1.5mg, Tab	12,690	30,106
Methyldopa 250mg, Tab	1,614,900	462,400
Metoclopramide HCl 10mg, Tab	740,500	175,750
Misoprostol 200mcg, Tab	255,728	82,984
Needle, Hypodermic, Luer, 21G, Sterile, Disposable	–	228,500
Needle, Hypodermic, Luer, 23G, Sterile, Disposable	1,557,800	922,050
Neomycin & Bacitracin 0.5% & 500IU/g, Ointment, 15g, Tube	53,390	19,975
Oral Rehydration Salts (ORS), Sachet	328,292	1,490,068
Oxytocin 10IU, Inj, Amp	194,025	170,970
Prednisolone 5mg, Tab	–	338,500
Progesterone-Only (Microlut) Levonorgestrel 30mcg, Tab, Cycle	–	254,967
Rapid test kit, Pregnancy, Pcs	152,075	68,625
Salbutamol 100mcg/dose, Aerosol, Inhaler	89,789	53,265
Sanitary Pads, Pcs	–	1,536
Sodium Chloride (Normal Saline) 0.9%, IV Inj, 500ml, Bot	61,559	4,747
Syringe, Luer, 10ml, Disposable, Pcs	1,535,300	805,950
Syringe, Luer, 20ml, Disposable, Pcs	709,840	200,880
Syringe, Luer, 2ml, Disposable, Pcs	1,308,600	802,400
Water for Injection 10ml, Inj, Amp	–	346,671
Zinc Sulphate 20mg, Dispersible, Tab	5,697,600	1,689,700

Note: – means no allocation during the time period.

Table S7: **Percentage of Missing Data by Product.** We report the percentages broken down between pre- and post- deployment of our system in 2023 Q2. The p-value denotes when differences are statistically significant.

	Pre-treated			Post-treated		
Product	Control	Treatment	p-value	Control	Treatment	p-value
Albendazole 400mg, Tab	0	0	1	1	0	1
Aluminium Hydroxide 500mg, Tab	98	99	0.422	95	92	0.157
Ampicillin 500mg, Pdr for IM/IV, Inj, Vial	15	25	0 ***	6	12	0.001 ***
Benzyl Benzoate 25%, Emulsion, 100ml, Bot	99	98	1	70	85	0 ***
Cannula, IV, 24G, Short, Sterile, Disposable, Pcs	40	29	0.001 ***	8	3	0.003 ***
Ceftriaxone 250mg, Pdr for Inj	96	99	0.014 **	89	88	0.907
Chlorhexidine Gluconate 7.1%, Gel	29	17	0 ***	4	5	0.654
Ciprofloxacin 500mg, Tab	90	89	0.392	68	20	0 ***
Erythromycin 125mg/5ml, Pdr for Susp, 100ml, Bot	100	99	0.64	83	85	0.363
Ferrous Sulphate 200mg, Tab	77	78	1	84	53	0 ***
Folic Acid 5mg, Tab	0	0	0.586	2	6	0 ***
Glove, Exam, Latex, Medium, Nonsterile, Disposable, Pcs	5	7	0.136	7	4	0.047 **
Glucose (Dextrose) Hypertonic 50%, IV Inj, 50ml, Bot	100	100	1	87	92	0.047 **
IV Giving Set, Pcs	5	4	0.419	5	5	0.889
Methyldopa 250mg, Tab	0	0	0.432	0	0	0.675
Metoclopramide HCl 10mg, Tab	90	88	0.339	86	83	0.361
Neomycin & Bacitracin 0.5% & 500IU/g, Ointment, 15g, Tube	98	99	0.572	71	68	0.527
Oral Rehydration Salts (ORS), Sachet	0	0	1	0	1	0.359
Oxytocin 10IU, Inj, Amp	2	2	0.723	4	4	0.67
Paracetamol (Acetaminophen) 250mg, Dispersible, Tab	0	0	0.589	0	0	1
Paracetamol (Acetaminophen) 500mg, Tab	37	30	0.041 **	12	22	0 ***
Povidone Iodine 10%, Solution, Bot	90	89	0.811	6	22	0 ***
Prednisolone 5mg, Tab	100	100	0.09	93	90	0.13
Salbutamol 100mcg/dose, Aerosol, Inhaler	88	79	0 ***	41	47	0.133
Water for Injection 10ml, Inj, Amp	27	26	0.889	33	32	0.818
Zinc Sulphate 20mg, Dispersible, Tab	0	0	1	0	0	0.639

Note: Significance levels: *** $p < 0.01$, ** $p < 0.05$