

邓科君合作研究依据与技术实施设计

本附件与导师沟通主稿配套，说明每个合作入口如何落实为数据结构、机器学习任务、研究比较与可复现产物。核心定位是由申请人牵头临床药学数据科学研究，既承担数据库与抽取工程，也承担算法建模、数据挖掘和科学问题解释，而不止提供病例或开展模型末端验证。

这里列出的模型、数据集、队列与查询均是拟议设计，尚无新增患者数据、模型训练结果或临床部署。方案中的工作量示例用于说明启动方式，不是已取得样本、统计效力依据或成果保证。研究问题可以共同选择，数据语义和评价纪律则应贯穿各入口。

阅读路径：第一部分核对已有证据和候选入口；第二部分给出数据资源与抽取工程；第三部分细化机器学习、数据挖掘及疾病任务；第四部分列出启动交付与证据使用边界。数据结构和任务可分别评审，不要求首期同时实施全部方向。

第一部分 已有基础与切入口比较

团队基础决定合作方式

邓老师及其合作的 LinDing 研究群已经具备计算方法、数据库、临床数据和疾病研究积累。网站列出的所有资源并不都由邓老师独立主持，因此本计划分别表述本人参与研究与合作研究群背景，不据网站论文汇总推定其当前项目安排。

电子科技大学邓科君主页、研究群合作人员页、数据库栏目、疾病研究栏目共同构成本计划的身份与研究版图依据。新任务的价值在于将既有药物表示、知识资源和临床研究连接到条件明确、时序准确、可计算且可审计的新研究对象，而不是重复建设一个泛药物网站。

直接承接的联合用药研究

DeepIDC。 原始任务为注射剂配伍参考表标签分类。原文基准集为 142 种注射剂、3396 药对；五类标签合并为三类，其中“单独用”和“稀释后可配伍”并入“不宜联合”。这些是参考表标签，不是患者不良反应；原文案例核对也混有体内药理警示，不能把全部旧标签视为纯体外相容真值。迁入新资源前，应逐来源重审证据域和结论粒度，不做直接改名导入。原文已经讨论溶剂、浓度等条件的限制；因此仅提“加入条件”不足以构成新意，应建立可复核的条件记录、来源隔离测试和具体条件下错误放行评价。其 warm/cold 任务为不同药物可见性的组合预测，不等于跨院患者验证。来源：原始论文 DOI 10.1007/s40262-022-01180-9，方法 2.1 至 2.4、结果与讨论。

MDDC。 该研究以配伍数据与 DrugBank 交集 2724 药对开展 DCI 和 DDI 双任务。DDI 由原文本模板类别归并为“药代动力学变化”和“不良作用”，并非“有或无 DDI”，更不是患者有无 ADE；DCI 仍为三类配伍标签。既有多任务与分子表示可作为基线，新增任务需要处理标签粒度、条件、来源依赖及跨域迁移，而非将多任务学习本身写作创新。来源：原始会议论文 DOI 10.1109/BIBM62325.2024.10822231，791 至 796 页正文和方法。

ACRM。 已核公开摘要与出版商预览显示，该研究使用三医院临床抗菌药共处方和有效性信息，建立推荐模型并报告独立回顾队列表现。因此不能把它描述为只有处方频次、没有临床数据或从未验证。完整正文及补充方法尚待合法取得，故本计划不推定其有效性操作定义、患者去重和具体分割方式，也不声称已经复现或发现其存在泄漏。新研究将另行冻结自身的患者—治疗阶段任务定义；若不能取得可运行原模型，则只比较可复现的基线，不冒称完成 ACRM 外部验证。来源：PubMed 38431108、原始出版记录 DOI 10.1016/j.ijantimicag.2024.107122。

慢病 数据资源与中药接口

团队 2023 糖尿病联合用药综述讨论获益、风险、个体差异和依从性，2026 并发症 AI 综述讨论外部验证与决策效用。这些是研究背景，不是新的患者干预效果。既有糖尿病/冠心病识别模型也不能直接回答“同一患者选择哪种治疗更有收益”。本计划重点增加治疗起点、实际暴露、共用药、停换药原因及重复决策过程。糖尿病联合治疗综述；并发症 AI 综述。

P450Rdb 整理酶催化反应与底物/产物，可支撑机制证据索引；它不是人体临床 DDI 结论库，物种、实验体系、剂量可达性与抑制/诱导证据必须另行限定。TCM2COVID 证明资源建设已有积累，不需借本项目重新建设 COVID 库。scRiskCell 的公共胰岛单细胞研究可以提供疾病机制背景，但细胞数不是独立患者数，也不直接提供临床用药反应或 ADE 标签。P450Rdb 原文；TCM2COVID 原文；scRiskCell 原文。

中药方向的团队思路已经涉及产品质量、临床分子轨迹、成分暴露及疗效分层。申请人的新增职责是把这些问题落实到可链接的产品/组方版本、真实暴露、共同治疗、表型与应答数据，并设计可证伪分析，

而不是再提出一遍“临床—基础—临床”。相关思路稿不写作已完成的临床试验，羊肚菌组学不写作糖尿病患者研究，AB4 衍生物的细胞/动物证据不外推人体疗效。2021 质量框架；2024 轨迹框架；2026 临床研究思路；2026 中尺度机制思路。

候选入口按可检验增量比较

入口	要回答的研究问题	申请人主导的核心产物	启动条件与降级路线
条件化药理学证据与抽取	条件和来源依赖能解释多少表面冲突 机器抽取能否可靠恢复这些条件	数据本体 证据库 抽取与实体链接模型 独立评测集	可先用合法公开材料启动 关键条件缺失时研究证据缺口
纵向药理学数据与暴露测量	医嘱 给药 接触与作用窗口的误差如何改变纳入与估计	CDM 扩展 ETL episode 算法 测量误差研究	需院内审批和源字段核查 缺给药记录则限制为处方任务
慢病多药动态风险与治疗应答	用药组合与病程变化何时增加可干预风险 或改变治疗相对收益	时序特征与组合挖掘 动态预测或独立因果分析包	需真实随访和可比策略 不能识别相对效应时保留预测
抗感染复评与组合迁移	组合知识在特定感染决策点能否增加患者级信息	病程队列 条件化组合模型 适用域与拒答模型	需明确感染场景与可见时间 结局不足时限定适宜性任务
中药产品 暴露与应答	产品差异和共同用药能否解释部分疗效异质性	产品组方版本模块 暴露应答模型 条件允许的分子增量分析	先有标准产品和可追踪用药 无样本则不承诺组学闭环

这不是五个同时启动的博士课题。可以先选择一项公开证据与方法任务，再与一项慢病或抗感染应用形成连续研究主线。二者共享数据与方法资产，但各自保留独立问题、终点和评价集。

竞争工作作为基线而不是否定方向

数据库方面，DDInter、Stabilis 和 OMOP 分别已有相互作用知识、配伍条件及临床标准模型。应比较本项目在选定主题内的条件完整性、来源去依赖、实体映射、真实暴露连接和可回答查询，而不能宣称此前没有相应数据库。

临床方法方面，INSPIRE 已评估患者特异抗菌用药提示，TriMaster 与后续治疗选择模型已研究降糖治疗差异。配伍方向也已有管路规则、预测工具比较与药师前瞻观察。由此，本项目的竞争性不在“首次 AI 加药师”或“首次个体化用药”，而在临床药理学测量问题、条件化数据与跨情境学习任务能否产生可重复的增量。下述评价应允许结果显示简单规则已足够，或复杂条件并未改善可迁移性。INSPIRE 原始试验；TriMaster 原始研究；2025 治疗选择模型；2023 配伍管路算法；2026 e-Harvis 研究；2026 药师 ICU 研究。这些来源的核验层次见附件末尾，不据摘要推定全文未见的细节。

第二部分 数据资源与抽取工程

条件化联合用药证据资源的实体与边界

核心实体与表

表/主键	最小字段	设计理由
ingredient / ingredient_id	标准通用名、中英文别名、结构/外部 ID、盐型关联、映射版本	活性成分相同不等于制剂相同；外部 ID 多对多并带来源
product / product_id	成分列表、成分含量、剂型、规格、厂家、批准/注册标识、辅料可知范围、来源国家、有效起止	中国本地制剂保留原码；不把无法匹配的药强行映射到美国品牌
source / source_id	DOI/PMID/注册号/说明书 setID、标题、版本/更新时间、获取日期、URL、许可层级、内容哈希、撤回/更正状态	同一论文摘要、数据库转载和补充表不是独立证据
study / study_id	设计、物种/体系、研究地点、样本单位、纳排条件、随访、试验/队列标识	区分同研究的多篇报告；患者数、实验孔数与记录数不混计
evidence_record / evidence_id	study_id、source_id、页/表/段定位、证据域、原始结论、结构化结论、抽取者、审校状态、版本	一条记录是一次可定位的“条件—观察/声明”，不是唯一药对
exposure_arm / arm_id	evidence_id、成分/制剂、角色、剂量、单位、途径、频次、疗程、顺序、比较组	联用为超边/方案，可含三种以上药；药对投影只是派生视图
context / context_id	年龄/人群、疾病、肝肾功能、感染部位、菌种/耐药、基线治疗、纳排边界	未报告 ≠ 正常；字段有 unknown/not_applicable/not_reported
endpoint / endpoint_id	终点域、定义、测量方法、窗口、效应量类型/值/区间、分母、单位、方向	RR、OR、HR、AUC 比值、实验粒径、HbA1c 变化不在同列强行比较
assertion / assertion_id	evidence_id、context_id、endpoint_id、证据类型、支持/反对/不确定、适用条件、审校意见	事实/标签与建议行动分表；有伤害报告不等于已证实因果
evidence_relation / relation_id	上下游记录 ID、关系类型、理由、审校者	引用、复述、共享队列、支持、条件不同、直接矛盾、被更正分别保存

外键全部指向本地稳定 ID，外部词表是映射表而非唯一真值。药对无方向视图可按成分 ID 排序形成 pair_key；PK 施加影响药/受影响药保留方向，不能因排序交换角色。多成分制剂既保留整体产品，也连接成分，不将复方的一条临床结论复制成所有成分两两因果关系。

四类证据分开 允许联系 禁止压成一个 安全 不安全 标签

迁入既有标签前逐来源重审证据域：DeepIDC 原始配伍表标签及案例核对混有体内药理警示，不把全部旧标签直接改名为纯体外 DCI 真值。旧类名、原始出处与新证据域并存，无法可靠分域的记录暂留未定，而非推断补齐。

· **A1 体外 DCI/稳定性**：product A/B、混合比例、各自与终浓度、溶媒、温度、接触时间、Y-site/同袋/注射器、容器与管材、检测方法、观察时点；物理观察、亚可见颗粒、含量/降解分别列终点。只测外观的“未见沉淀”不能升格为化学稳定。无实验条件不能自动作为条件真值。

- **A2 体内 PK/PD 相互作用：**施加影响药/受影响药、剂量顺序、稳态/单剂量、受试人群、酶/转运体、AUC 或 Cmax 比值与置信区间、临床 PD 指标、体外/健康志愿者/患者证据标记。P450 催化底物关系不直接等于抑制、诱导或临床 DDI。
- **A3 患者安全：**队列/试验或病例/自发报告来源、药物角色、事件定义/时间、对照、分母是否已知、归因过程。自发报告保留信号与不确定性，不产生发生率标签；“潜在 DDI 警报”与“实际 ADE”不同表。
- **A4 疗效：**适应证、治疗阶段、对照策略、共同治疗、主要/次要结局、观测窗口、失访与效应量。体外协同分数、处方频次、药师认为适宜和临床疗效分别编码。

这样同一组合可以同时是“某条件不能同管接触”“临床分途径治疗有适应证”“未发现特定 PK 影响”“患者事件证据不足”，不构成数据库内部自相矛盾。

一条真实来源的结构化事实演示

来源为 DailyMed Baxter 头孢曲松注射液说明书，setID=4c5c2d3f-5038-41a1-a2fe-4dcd048dbac1，页面更新 2024-10-07；核查 2026-10-24。原说明书。下表是**来源事实抽取示范**，不是新的实验数据，也不能替代中国具体制剂说明书。

字段	值
evidence_id	DEMO-LABEL-CTX-CA-001（本稿演示 ID）
source_type / evidence_type	药品说明书 / 条件性给药限制声明
source_locator	WARNINGS → Interaction with Calcium-Containing Products
ingredients	ceftriaxone; calcium-containing IV solutions（类概念，不伪造单一钙制剂 ID）
condition	同时静脉给药、包括 Y-site；需保留新生儿与非新生儿区别
assertion	说明书不允许上述同时给药；非新生儿可在符合冲管条件下顺序给药
condition_qualifier	两次输注之间以相容液体充分冲洗管路；不将“分两条管路”或“错开一会儿”擅自写成等价条件
domain	条件性静脉配伍/给药安全声明；不是定量 PK 效应量
concentration / contact_time / experiment_n	本条声明未提供；not_reported，不从其他章节猜填
patient_ADE / causal_effect	未在本条构造患者事件或效应量
normalization_status	名称已人工核查；跨国产品可迁移性未确认

真正的数据工程增量：不能把这段说明书压成 ceftriaxone+calcium=1。应拆出“同时给药限制”与“非新生儿、满足冲管的顺序给药条件”两个关联声明，并关联说明书 CONTRAINDICATIONS 中≤28 日新生儿的独立限制：需要或预期需要静脉含钙制品治疗时禁用，不能把非新生儿顺序给药条款套用过去。避免用机器自动生成的泛化替代完整标签核查。

资源 B 以真实用药过程为中心的纵向药学 CDM

基座复用 不重造 OMOP

以 OMOP CDM v5.4 的 person、visit、drug_exposure、measurement、condition 等为跨研究交换基座；保留本地药学扩展表和原始值。OMOP 已有处方、发药、给药来源类型区分，因此本方案不

是“首次提出药物暴露标准”。增量是本研究所需的执行级时间、制剂/配液/管路、治疗策略版本、停止原因和证据桥接。OHDSI 规范。

表与主外键设计

实体/主键	核心字段与关系
person / person_key	院内可信域生成的研究假名；出生信息按审批最小化；重复就诊关联不靠姓名导出
visit / visit_key	person_key、入出院/门诊时间、院区病区转移、科室、观察覆盖范围
observation_window / window_id	person_key、来源系统、可观测起止、门诊/住院覆盖；院外未知不能写无事件
med_order / order_id	person_key、visit_key、product_id、开立/计划起止、途径、剂量频次、取消状态、医嘱版本
dispense / dispense_id	order_id、实际发药/退药时间、数量；只证明药物提供，不证明服用
administration / admin_id	order_id、actual_start/end、实际剂量/单位、执行/拒绝/漏服/暂停、记录时间、来源记录 ID
preparation / prep_id	对应 admin_id、制剂批次可得性、溶媒/体积/浓度、配制时间、静配系统记录
line_contact / contact_id	两个或多个 admin_id、lumen_id、接触类型、共同接触区间、冲管及来源可信度；未知单列
measurement / measurement_id	person_key、visit_key、项目/原值/标准值、单位、标本采集时间、结果可见时间、异常标记
microbiology / micro_id	标本部位/采集时点、病原、鉴定/药敏结果及发布时间、MIC/解释、检测标准版本可得性
clinical_state / state_id	诊断和发生/记录时间、感染部位/严重程度依据、肝肾状态、合并症、来源与算法版本
regimen_episode / episode_id	person_key、visit_key、策略起止、目的、成员 admin/order ID、剂量变更、停换原因、构建规则版本
decision_point / decision_id	episode_id、研究任务、index_time、当时已知信息截止、候选策略定义、纳排规则版本
outcome / outcome_id	person_key、事件时间/判定窗口、定义版本、客观证据 ID、人工裁定、竞争事件/随访状态；疗效安全分域
pharmacist_action / action_id	decision_id、问题类别、证据 ID、建议/接纳/实际执行时点、结果；“接受建议”不写成疗效
evidence_match / match_id	decision/admin/contact ID、A 的 assertion_id、匹配/不匹配/未知字段、规则版本、依据日期

每个事件至少保留两个时间：**发生/采集时间和记录/结果可得时间**。若结果在决策后才回报，不能作为决策当时特征。第三个时间是研究仓库入库时间，用于回放、版本和质量追踪。药物暴露窗口的端点采用统一约定（例如左闭右开），在构建规则中写明；一条事件重复入库、订单更新与实际重新给药必须区别。每次研究切片还记录抽取截止、队列定义版本与来源观察截止，不能以今天更新的记录无标识地重建过去决策。

暴露不能只有一种 四层并存

1. co_prescribed：医嘱区间重叠，只表示同期开立。

2. co_administered: 真实执行区间或经预设给药后作用窗口重叠; 口服门诊只能按发药/自报等证据定义“可推定暴露”, 不能冒充已服药。

3. co_contact: 同一管腔/混合容器且接触区间重叠, 需要管路或配液依据。不能从同日给药推定。

4. interaction_risk_window: 按研究中的 PK/PD 问题预先定义残留/诱导等作用窗口, 可能与输注时间不同。不得把“用药已停”一律等于作用消失。

精确执行时间不可得时保留区间不确定性; 不把 00:00 默认值误当真实时点。所有推断都有 observed/inferred/unknown 及算法版本。

完全虚构的患者时间线 演示数据模型 不是临床处置建议

P_SYN_001, 虚构成人、虚构住院 VIS_SYN_001, 以下全部临床数值与事件均为合成。用“住院第几日”避免制造真实病历外观。

时间	原始事件	数据模型如何表达
入院前 30 日	门诊处方长期慢病药 X; 实际服用无记录	order 存在; 依从性 unknown; 不能连续补齐 30 天已服用
第 1 日 08:00	住院, 临床记录慢病合并感染	visit 开始; 感染表型候选, 保留诊断依据和记录时间
第 1 日 08:30/10:00	肌酐标本采集/结果可得, 合成值 90 $\mu\text{mol/L}$	两个时间; 09:00 决策不得使用 10:00 才可见结果
第 1 日 09:00	开立抗菌药 A 与 B 医嘱	co_prescribed=是; 实际给药未知
第 1 日 10:30–11:00	A 已执行; B 医嘱 10:40 取消, 未执行	A admin 有; B 只有取消 order, 不能生成 A+B 实际暴露
第 2 日 09:00–09:45	A 执行; B 09:20–10:00 执行, 管路字段缺失	有给药时间重叠 25 分钟, 但 co_contact=未知, 不能判定配伍接触
第 2 日 11:00	前一日培养结果此时回报	只能进入 11:00 后的决策信息集
第 3 日 08:00/09:00	肌酐采集/可见, 合成值 150 $\mu\text{mol/L}$	肾功能变化候选事件; 是否符合预设终点须规则/裁定, 不能自动归因 DDI
第 3 日 10:00	方案调整, 记录原因为“病情/药敏复评”	新 regimen episode; 停换原因为记录事实, 不偷换为药物不良反应
第 7 日	出院后无连续院外随访	住院观察结束; 院外事件 unknown, 不能填随访 30 日无 ADE

这条时间线能显示申请人的贡献: 清除“医嘱共现即联用”的误分类; 保留未知管路; 阻止未来检验泄漏; 使停药原因、结局和药物因果判断可分别研究。

数据从哪里来 什么可以公开 什么只能院内做

层级	来源与首期用途	使用/发布边界
P0 公开元数据	PubMed/DOI 元数据、试验注册条目、公开资源目录; 建立来源索引	注册不等于试验已完成, 结果未公布 $\neq$ 无效; 许可与接口条款逐源登记
P1 可合法使用的全文/补充表/说明书	开放原始研究、授权获取的文献、官方说明书; 抽取条件/终点/证据	开放可读不等于可再分发全文; 发布规范化事实、索引和短定位, 版权文本/图表另审

层级	来源与首期用途	使用/发布边界
P2 第三方数据库	DeepIDC/MDDC、DDInter、P450Rdb、TCM2COVID 等合法可用内容	数据、代码、模型许可分别核查；可只链接/映射，不复制受限内容；不能承诺其完整下载已完成
P3 资格受控研究数据	MIMIC 等可作跨系统 ETL 与表型压力测试备选	必须个人授权/培训/DUA；需在获得资格后申请和使用；不能拿来替代中国慢病连续门诊队列或真实输注管路证据
P4 院内受控数据	经批准后 HIS/EMR、LIS、微生物、静配、护理执行、药师记录、必要随访	院内可信计算、最小权限、审计；默认不向云端上传真实记录；不因去标识就视为可公开
P5 新增前瞻记录/样本	后续有明确问题时记录药师裁定、依从性、管路，或加组学	另行审批、方案、知情同意/豁免评估；不是本次计划已拥有的资源

FAERS 等自发报告若纳入，单独保存报告版本、药物角色、事件、重复处理和信号定义，仅用于候选关系/证据覆盖，不与医院病人按药名或年龄“拼接成同一个人”。FDA 强调公开安全报告有重复及字段限制，不能据此直接比较药品安全性。FDA 说明。

MIMIC 权限需逐人、逐数据集遵守，不能由一个人获取后分发全组；在线服务政策也须按当时条款和机构批准核验，不预设允许外发。PhysioNet FAQ、MIMIC 获取说明。

数据提取与可复现研究资产

文献 证据 ETL

- 先设计检索与数据字典。** 明确药物范围、四域纳入规则、语言、文献类型；用种子药对/团队论文引用链做试采样，保留检索式、日期与排除理由。范围可扩，但每版有封闭边界。
- 来源登记与合法归档。** 原始文件只读保存、checksum、获取途径/许可、DOI/版本；用标题/DOI/试验号查重。OCR 仅用于确有需要的扫描文档，并记录 OCR 质量，不因文件名正确就认定内容正确——ACRM 文件名与正文错配说明必须按内容识别来源。
- 版面与证据定位。** 表格行、脚注、单位、否定句、限定词、各臂样本量、补充表一并解析。输出字段值必须关联页/表/原片段；无法定位的机器推断不得进入事实层。
- 实体规范化。** 中文通用名/商品名/盐型→本地药典；再映射国际词表。精确规则优先，模糊匹配只生成候选；一对多、复方、相同名不同制剂入人工队列。保留原值和映射置信/人工状态。
- 条件与数值规范化。** 原值、标准值、单位、换算公式共存；浓度要说明成分基准/盐基准/最终混合基准。无法换算保留 unknown，不能猜密度/体积。不同温度/时间分成不同实验条件，不用平均值吞掉差异。
- 机器辅助抽取作为方法研究。** 规则/词典→轻量文本模型或获准工具的受约束结构化抽取→校验器→人工裁定。模型只能提交候选，不能自造未报告字段；困难样例、中文缩写、表格跨页和否定条件单列测试。无需先下载大模型证明能力。
- 去重与来源依赖。** 文档重复、同实验多次报道、同患者队列后续论文分别处理。记录共享实验/队列 study_family_id，不按“多个网站同说法”计为多条独立证据；用条件指纹找近重复，冲突保留而非多数投票删除。
- 弱标注与人工真值分层。** 数据库标签/说明书规则为弱标签，人工审校为“抽取正确性”金标准；实验兼容真值只能来自合格实验。患者事件以预设、盲法病历裁定形成临床参考标准；它仍可能受病历缺失与专家分歧影响，不等于 DDI 因果金标准，也不与药师认为需复核混用。
- 双人独立标注、第三人裁决。** 先小样本试标、修订手册，再冻结测试集；高风险和冲突记录优先复核，但还要概率抽样估计总体误差。不能只用主动学习选中的难例计算全库准确率。

10. 版本发布。schema_version、vocabulary_version、source_snapshot、extraction_version、annotation_version、release_id 六项随包发布；新增、更正、撤回标记有 diff；旧研究锁定版本可再运行。

医院数据 ETL 与隐私安全连接架构

医院源系统（HIS/EMR/LIS/护理/静配）

→ 院内只读接入与数据合同（经审批的表/字段/用途）

→ 原始受控区（可回溯 source_record_id，不向研究者开放身份）

→ 院内身份关联服务（生成项目级 person_key、visit_key）

→ 标准化区（词表、单位、时点、取消/退药/执行状态）

→ 质量隔离区（不合理剂量、重复、时间倒置、映射失败）

→ 药学扩展 CDM + 可回放暴露/结局规则

→ 冻结研究数据集（任务专用最小字段）

→ 经审批的院内分析 / 跨机构代码到数据

→ 人工披露审查后的汇总、数据字典、代码、合成示例

- 姓名、证件、手机号等在可信身份区；研究者拿不可逆项目标识并不等于全数据已匿名，仍按敏感研究数据管理。**不把直接 SHA256 姓名/身份证当可靠隐私保护**；标识生成和密钥由医院受控服务负责。
- 默认仅院内确定性关联。多院数据先采用统一字典与本地执行、汇总对照；若确需跨院识别人，另立合法依据与安全链接方案，不能靠低熵哈希或人口学相似度偷偷连接。联邦分析也有信息泄漏风险，不自动视作合规。
- 提取采用增量水位/变更版本与幂等主键；迟到记录可重算受影响的 episode 而非重复追加。订单修订、取消、退款不覆盖执行原始事实。
- 权限按任务最小化、读写分离，访问与导出审计；保留期、撤销访问、研究结束处理写入方案。文本脱敏需抽检残留标识；公开只给合成文本与抽取代码。
- 数据按来源/日期分区，优先 SQL 与列式存储完成联结、窗口运算、审计；达到单机资源阈值再考虑分布式处理。**大数据能力体现为增量、版本、性能与错误控制，不是预先堆平台名。**
- 每个自动化任务须有输入合同、唯一键、行数/去重/单位/时间断言、失败隔离与重跑记录；试运行比较“输入记录数→排除/修订/隔离数→输出数”实现核对。先用合成数据做边界测试，再在获批院内数据上做源记录回查。日常增量与研究冻结 release 分开，避免业务更新改变已发表结果。
- 许可登记至少拆为“允许访问、允许本项目分析、允许派生数据发布、允许全文/原表再分发、署名义务、期限/撤回条件”，逐项保留原条款链接与审核状态；未明确的发布权限视为待核，不因可下载就允许公开。词表本身也需核查许可，尤其不能把受限术语全量附带到公共包。

五项目复现研究查询

这些是计划支持的查询，不是已在真实数据库运行的结果。每项都可先用合成/公开小样例展示查询合同和预期输出。

查询	所需表/条件	输出与科学意义
Q1: 同一药对的配伍“冲突”有多少由条件不同解释?	A1; 同 pair_key 分组, 再比制剂、浓度、溶媒、温度、接触方式/时间、检测方法; 按 study_family 去依赖	条件可解释冲突、真正同条件矛盾、信息缺失三类及逐条来源; 为 DeepIDC/MDDC 构造条件化任务, 而非直接合并多数标签
Q2: 抗菌“医嘱联用”有多少实际执行、多少能确认同管接触?	B 的 orders→administrations→line_contacts; 按住院 episode 和 drug pair 统计	三层分母、误分类方向、病区/药物差异; 量化训练标签和研究暴露测量误差, 不把未知算否定
Q3: 慢病患者开始抗感染前后, 哪些联合用药风险证据适用性发生变化?	固定 decision_point, 连接当时可见肾功能/病原信息、长期药物暴露证据、A2/A3 适用条件	适用/不适用/缺关键条件的清单及理由; 可支撑之后的动态学习任务, 不输出自动停药指令
Q4: P450 机制中哪些人体相关关系有临床 PK 证据、哪些仅体外线索?	P450Rdb 合法映射→物种/酶→A2 的试验体系、剂量/浓度、方向与效应量→A3 事件证据	“反应—机制—PK—安全”分层证据路径; 统计跨层证据断点与冲突, 不能把图路径长度当因果强度
Q5: 同一慢病治疗策略下, 能否区分疗效差与暴露不足/停药不耐受?	B 的启动 episode、医嘱/发药/执行、自报、终点窗口、停药原因; A4 适用人群	形成可复现疗效分析队列和暴露质量分层; 报告可观测分母和未知比例, 支持治疗应答/异质性研究

第六个可选查询是：同名中药制剂/方剂的产品与组成差异能解释多少“同药不同效”表象？先有真实产品和组方数据，再连患者/样本；不根据药名将不同来源制剂当同一干预。

竞争资源与可检验增量

基线	已有的能力	本方案不能宣称什么	值得检验的增量
DDInter 2.0	药物相互作用、机制、风险分层、管理信息及处方检查等；官方站可查	“首个药师整理 DDI 库/首个含管理建议图谱”	具体研究范围内逐原始证据的条件与版本、来源依赖、制剂及真实暴露连接；需要头对头覆盖与正确性比较
Stabilis	稳定性、配伍、Y-site 表、制剂及参考文献	“此前数据库只有药名，没有配伍条件”	不同来源和本地制剂条件映射质量；未报告/不可外推的显式表示；条件冲突归因及基准复用
OMOP/OHDSI	标准临床数据模型与多类药物暴露来源	“首个统一医院数据/首个纵向药物 CDM”	药学接触拓扑、给药—决策可见时间、停换原因、分层证据桥、任务级测量误差研究
DeepIDC/MDDC/ACRM	团队已开展药物组合算法、不同标签域和真实临床研究	“团队没有数据库/没有临床/没有多任务”	延续既有能力，建设条件化、来源隔离、患者时序的新数据问题及可复用基准

基线	已有的能力	本方案不能宣称什么	值得检验的增量
P450Rdb/ TCM2COVID	已有机制/中药资源建设经验	“重新搭网站就是创新”	实体对齐、适用域与证据链的临床化扩展；有明确问题时链接而非复制全库

来源：DDInter 官网、Stabilis 官网、OMOP v5.4。DrugComb 的当前版本和字段尚未核实；若进入体外协同分支，再补原始资源论文与数据字典。数据库存在说明可复用与可比较，并不证明本研究无新意；新意要用新增记录/条件、纠错、可计算查询与跨源复用的证据证明，而不是宣称全球首次。

从工程产物到研究论文 独立科学贡献怎么成立

资源 A 数据论文 提取 规范化方法论文

可检验假设 A：相较药对级聚合，显式保留条件、证据域和来源依赖，能降低“伪冲突/伪独立证据”、提高特定问题的可回答性，但可能暴露大量证据不足；结果可能支持条件化的必要性，也可能显示某些范围无需复杂建模。

- 数据论文：边界明确的主题资源、原始证据定位、版本发布、许可矩阵、可查询字典和真实/合成演示、覆盖偏倚分析。
- 方法论文：抽取中的联合实体—关系—数值/单位—条件约束、跨语言药名/制剂映射、矛盾与来源依赖识别；与规则、通用抽取、无条件聚合比较。不是用 LLM 代替人工后只报“省时间”。
- 评价：字段级 precision/recall/F1 及分层置信区间；数值/单位完全匹配；否定/条件遗漏率；制剂误合并率；来源定位可复核率；冲突分类一致性；每条经验证记录的人时/成本。误把禁忌读成可用等严重错误单独计数。
- 金标准按来源类型、中文/英文、表格/正文、药物/证据域分层；测试集先冻结，抽取规则调优后不回灌。按研究家族/来源和时间隔离，防止同一实验转载跨集。

资源 B 药学表型 暴露测量方法论文 队列资源论文

可检验假设 B：医嘱共现、给药重叠、同管接触与药理风险窗口的误分类程度不同，且对联用研究的纳入人群、事件分母和效应估计产生可量化影响。

- 与盲法人工病历/执行核对相比，评估暴露 episode 和停止原因的敏感度、阳性预测值、时间误差、边界一致性；每类评价都说明审核者可见材料和真值限制。
- 比较“粗医嘱算法—执行算法—条件/时序算法”的差异；明确新增字段可得性与维护成本，不能只宣称精细一定更好。
- 队列质量：映射覆盖、单位一致性、缺失机制、时间完整、观察覆盖、重复住院关联、来源变化；按病区、年份、产品和患者亚组展示，不只一个总正确率。
- 跨院阶段先测 ETL/表型可迁移性与字典覆盖。不能因两院都转成 OMOP 就宣称临床可比。
- 开放内容可为字典、ETL/队列定义、质量规则、标注手册、合成记录、经审查统计；不以公开患者行级数据作为论文成立的先决条件。

两库一桥 综合论文

可形成“临床药学知识与纵向用药暴露的条件对齐框架”：以 Q1-Q5 中的两个代表查询评估，比较仅药名连接、仅药对标签连接、完整条件/时间连接的事实匹配正确率、可判断覆盖率、未知原因结构与复核工作量。这里先证明数据语义和证据匹配可用，再与算法模块合作研究预测增益。不能用引用检索成功率冒充临床决策有效性。

资源/方法/疾病应用是可衔接成果路径，不承诺每一个都会成文，更不预定期刊或论文数量。只有独立问题、足够数据质量和合适比较成立时才分篇，避免重复切割。

中药产品与暴露扩展表

该扩展只有在团队与医院实际使用场景匹配时开启，不把消渴方、消渴丸、玉泉丸或不同厂家同名产品互相替代。

表	字段与链接
formula_version / formula_version_id	方名、经典/医院/商业来源、组成、各药材剂量、炮制、加减、版本；每次加减作为新版本，不静默覆盖
herbal_product / herbal_product_id	批准/备案标识、厂家、剂型、规格、批号可得性、formula_version_id；每个批次是否有实测指纹单独标记
quality_assay / assay_id	产品/批次、检测方法、指纹峰/标志成分、单位、定量下限、批内/批间质控；缺测不以公开文献成分替代
herbal_exposure / exposure_id	person_key、产品/组方版本、实际剂量与频次、开始停止、给药/发药/自报证据、合并化药、依从性来源
phenotype_visit / phenotype_visit_id	预设慢病表型、标准临床指标、症状量表；若纳入证候需版本化定义、评定者和一致性，不以主观标签直接替代疾病严重程度
biosample / sample_id	person_key、采样事件、与最近一次给药间隔、空腹状态、处理/保存、组学批次、知情同意范围；组织类型不互换
response / response_id	预先选定疗效/安全终点、基线和评估时间窗、共同干预、停药失访原因，连接原始测量证据

核心科学用途：先分清“产品不一致、实际暴露不同、共同治疗变化、疾病阶段差异”这些可测因素，再研究组学/分子应答；不能把数据库列出的全部成分都视为患者体内暴露。血液代谢组可以形成独立研究，但不能声称替代团队胰岛单细胞数据。来源依据为已读 2021 中药质量框架及 2024/2026 思路稿，属于本稿的工程化设计推导，不是论文已经完成的数据库。

第三部分 算法 数据挖掘与疾病研究任务

首先锁定语义 不用一个安全分数包办所有问题

DeepIDC 的核心是注射剂配伍参考表/IDC 标签任务；其原始五类合并到三类的过程应保留，不能把“稀释后可配伍”或“单独使用”直接解释成患者发生不良事件。旧标签并非已经逐条分清证据域的纯体外兼容真值，原文案例核查也混有体内药理警示。因此资源迁移第一步是逐条重审标签出处、条件与证据域；无法回溯者保留“旧 IDC 标签/语义未核定”，不直接改名为新的 DCI 金标准。MDDC 在交集药对上联学 DCI 与 DDI，但其 DDI 二分类是药代动力学变化与不良作用类别，不是有无 DDI，更不是患者有无 ADE。ACRM 已经利用多中心临床共处方与有效性信息并报告回顾验证；其完整有效性定义、分割细节和补充方法目前仍待合法全文核清，不据错名本地 PDF 补齐。

本项目保留四个标签空间：体外物理化学配伍、体内 PK/PD 证据、患者安全事件、临床疗效。另设“药师可行动问题”标签，它描述复核/处理需要，不冒充前四类真值。原始标签、标准化标签、模型训练标签及裁定标签分别保存版本。一个药对可在某种混合条件下不相容，也可在分途径、不同患者条件下有治疗价值；这不是模型必须消除的矛盾。

标签中的未知也要拆开：没有检索到证据、证据存在但条件不匹配、来源相互冲突、患者暴露不确定、结局不可观察、无法裁定，不能统统编码为 0。预测概率、证据可信度、适用域和建议复核优先级是不同输出，不能以单一神经网络分数替代。

可执行的任务合同

下表给出首轮设计合同。窗口为说明如何落地的候选值，须在药物、疾病和数据质量确定后预设，并在任何锁定测试前冻结；不是临床规范阈值。

任务	分析单位与时间起点	输入窗口和可见性	标签与主要用途
T0 条件证据抽取	一篇研究中的一个可定位证据记录及其来源家族；无患者时间零点	固定版本全文、表格、脚注与合法补充材料	药物/制剂、剂量单位、条件、否定、终点、效应量及原文定位；建设可审计训练数据
T1 条件化配伍与 DDI 知识	药对或多药方案加具体条件；无患者时间零点	制剂、浓度、溶媒、接触方式/时间及研究体系；PK 角色保持方向	DCI 各实验终点/证据结论；PK 与临床 DDI 另立任务；预测证据适用与未知，不输出患者疗效
T2 抗感染可行动风险	患者的一次预定义复评决策点；例如启动后 48 至 72 小时的复评，而不是任意处方行	决策前病史与既往暴露，近 24 小时状态及截至该点已回报微生物；具体范围后定	基线信息下独立裁定的可行动问题；另评未来 7 日等固定窗口的客观安全/疗效结局
T3 慢病多药风险与应答	一次确切的加药/换药/减药决策，或一次预定义慢病合并感染阶段	例如前 90 日可观察病史及治疗前最近有效检验；仅用决策时已知信息	固定随访的连续疗效、独立安全事件、持续用药与停换原因；不先混成总分
T4 比较策略与获益异质性	在相同资格与时间零点下可接受策略 A 或 B 的患者	治疗前混杂因素与用药证据；动态问题另需完整纵向状态	明确人群、策略、随访、结局和估计量；是独立因果研究合同，不由 T2/T3 预测自动得出

每个合同还必须写明：目标人群、排除规则、数据版本、可观察范围、标签分母、主要与次要指标、可接受行动范围、模型服务对象和失败后的降级研究。多次住院、多个决策点、药对、给药记录与患者数分别统计，不能把记录量当样本独立性。

T0 从文献抽取到可训练标签的方法研究

输入与金标准

从资源 A 中选定一个封闭主题，保留文献获取、重复/共享队列关系和来源时间。标注单元为“证据片段及其上下文”，字段包括成分、制剂、施加影响药/受影响药、剂量、浓度、单位、给药方式、实验条件、终点、否定与限定词。将实体识别、实体链接、关系抽取、数值规范化、条件归属与证据定位分别评分，防止只报总体 F1 掩盖关键错误。

金标准先形成双人独立标注和第三人争议裁定手册，按中文/英文、表格/正文、证据域与来源类型分层。冻结测试来源家族，同一实验在论文、综述和数据库的复述不能跨训练测试。可建立困难挑战集，例如“未见可见沉淀但未测降解”“药物停用但残余作用仍可能存在”“降低某药暴露而非提高”等；挑战集用于压力测试，不能冒充真实分布性能。

可由申请人实际实施的基线

从规则与词典开始：药名/盐型/制剂别名库、单位模式、否定词和关系方向规则；再比较传统序列标注或文本分类、可复用领域编码器、合法可用的大模型结构化抽取。加入 schema 约束、单位一致性和跨字段校验，要求输出可定位来源，不允许无证据补齐数值。比较“先实体后关系”与“联合条件记录”方案，测量条件被错误挂接到另一药物或另一试验臂的错误。

采用人机协同而非全量人工或全量自动：用置信度、模型分歧、稀有制剂、否定句及高危字段错误作为主动学习候选，保留随机抽样审计，避免只审核疑难样本后误估总体精度。报告每条经验证记录的人时、成本、修订率和严重误读率；若自动抽取减少工时但提高危险条件遗漏，结论不能称改进。

弱监督不等于伪造金标准

药名词典、说明书规则、不同数据库声明、文本模型可以作为标注函数。每个函数具有适用域、来源、投票/弃权 and 版本；同一说明书被多库转载时并非多张独立选票。比较简单多数票、相关性可审计的弱标签模型与小金标准监督模型。训练弱标签与独立测试裁定隔离，使用人工金标准估计标签偏差，不在测试集上调整规则。

若某关系只有已报告阳性、其余均未标记，可探索 PU 学习，但必须说明标记机制。临床上更容易报告严重事件、常用药更容易研究，通常不能无条件假定已标记阳性是随机抽样。可做类别先验与选择机制敏感性分析，并优先补充独立阴性/不支持证据。没有明确阴性集时不宣称无偏特异度，不把 PU 分数直接叫真实发生概率。

T1 条件化联合用药模型与不确定性分流

输入及建模顺序

先比较现有资料检索与显式条件匹配规则，再比较仅药对特征模型、药对加结构化条件模型。简单基线可包括逻辑回归/多项回归、线性或核方法、随机森林、梯度提升；深度双视图或图模型须在强基线后验证增量。制剂、溶媒、浓度、接触方式、温度与时间不能因缺失而由模型任意补成“常见条件”。按实验方法分别输出外观、颗粒、含量/降解等终点；只有外观记录时不能训练“全面稳定”标签。

需要处理同药对多条件时，可用分层模型或条件编码，将药物表征、产品表征和条件分别建模，比较是否比简单拼接更能解释跨条件变化。多任务只连接有科学关系且标签可用的任务；对缺标签设置掩码，不将缺标签补负例。DCI 与 DDI 共享表示的收益要逐任务报告，检查负迁移；不能因总损失下降就认定临床更安全。

药对可为无向键，但 PK 角色、给药顺序与时间依赖须保留方向。三药以上组合采用方案/超边表示，不能把双药结论直接外推到三药混合。首期若三药数据稀少，应报告不可回答范围，而不是训练过度参数化模型。

优先评价什么

除多分类 macro-F1、逐类灵敏度/PPV 和混淆矩阵外，单列“在预设参考标准明确判为该条件下不相容或不宜接触的记录中，模型判可相容的比例”及置信区间；另列“在参考证据不足的记录中，模型仍肯定判可相容”的无证据肯定率。两者分别以相应已裁定条件记录为分母，是离线标签错误指标，不是临床实际放行或患者伤害发生率。对条件缺失或新药/新制剂设弃权：报告覆盖率、覆盖范围内错误风险、风险—覆盖曲线和拒答原因。比较在相同覆盖率、相同药师工时或相同允许危险误判上限下各方法的收益，不能通过拒答大多数困难样本制造高准确率。

药师的结论是证据是否支持、是否需复核及操作流程是否有问题，不是新的化学真值。对高频且证据空白的具体条件，模型只可排序；有条件平台的独立配伍实验才可补实验真值。没有此能力时，该任务仍可产出条件化资源、抽取方法和证据适用域研究，但不宣称新组合已经证实相容。

T2 抗感染病程级研究包

场景与暴露

选择一个确有可比复评决策的感染场景，先界定感染部位、治疗阶段、病原证据与严重度可得性，不把所有感染放进一个排行榜。主要分析可先每个感染 episode 选一个复评点；若后续做每日 landmark，必须说明每次风险集、重复事件和患者内相关性，不能把每日记录当独立患者扩充样本。

联用暴露由实际给药优先构建，医嘱、发药、给药、同管接触与药理作用窗口分别标记。模型特征包括基线合并症、既往抗菌药暴露、当前剂量与肝肾功能匹配、已知感染状态、已回报药敏、治疗阶段、其他可能影响毒性或疗效的合用药，以及条件化证据匹配结果。仅在决策后回报的培养、后续病情改善、停换药原因和出院诊断，不能进入该点的预测特征。

两类标签不可混成一个效果

第一类是**当前可行动药学问题**：适应证/覆盖、剂量或给药问题、重复或不必要暴露、需监测风险、证据不适用等可采用多标签编码，再独立设置紧急复核、常规复核/监测、未识别问题与信息不足。评审者只看时间零点前的信息，且对模型结果盲法。未识别问题不代表绝对安全，信息不足不当阴性。

第二类是**未来客观结局**：例如有可靠基线与检验覆盖的肾功能事件，或预先界定的感染疗效结局。阈值、基线取值、观察窗口、重复测量、死亡/出院/失访处理须由临床与统计共同冻结；本稿不代写临床诊断标准。事件判定可看完整随访材料，但这个包与基线适宜性评审包隔离。感染导致的器官损伤不能因同时用药就归为 DDI；停药、升级、降阶梯和医生接受提示不能单独当疗效或药物伤害真值。

模型与消融

建立递进基线：现行药师/抗菌管理规则；仅药对推荐或知识库风险；仅临床变量的正则化回归或 GAM；梯度提升等表格模型；临床加实际暴露；最后加入条件化证据与药对表示。根据样本与时序质量再考虑序列模型，不以网络复杂度作为创新。

关键消融要回答：增益来自真实给药还是简单药数？来自患者病情还是药对表示？来自已知药物标签还是条件信息？将医嘱共现、执行暴露和药理窗口分别构建同一队列，量化标签与性能变化。未经授权或没有原模型代码/权重时，只能比较可复现基线，不把重做近似模型称 ACRM 正式外部验证。

评估 AUROC 之外以 PR-AUC、固定复核容量下灵敏度、PPV、每百决策点告警数、严重漏报和校准为重点。反映工作量的指标必须记录实际复核时间；决策曲线和离线净效用依赖阈值/权重，作为辅助，不代替真实患者获益。先做锁定离线和前瞻静默验证，不自主改变处方。

T3 慢病多药的预测与疗效异质性研究包

两种可选但不应混做的入口

入口 A 以慢病合并感染阶段为桥，预测短期安全或需要药师复核的风险，强调肾功能变化、短期抗感染暴露与长期药物之间的关系。它能复用住院执行数据，但不能据此承诺长期心血管获益。

入口 B 选择一种明确慢病的加药/调整决策，例如糖尿病复杂共病患者中临床确有选择余地的策略 A/B，研究固定随访的疗效、安全与治疗持续性。必须有治疗前指标、既有方案、剂量、发药或服用证据、复诊与结局覆盖；只有横断面体检或住院收费记录时不成立。团队已有糖尿病/冠心病分类与治疗综述，因此不是再做一个既有疾病分类器。

标签与特征

疗效可选固定窗口的连续临床指标、基线调整后的变化或预定义达标；安全终点独立；不耐受停药、药效不佳换药、费用/获取困难、依从性不足和原因不明分别编码。没有记录的停药原因保留未知，不由算法从后续处方消失自动判断。院外处方与实际服用不同，有验证子样本才可评估其误分类；发药覆盖或自报依从性是代理，不是生物学暴露真值。

多药信息可包括长期药物类别、剂量、时间重叠、器官功能依赖、用药复杂度、近期新增/停止和风险机制类别。先比较简单药物数量与临床可解释的暴露信息，避免把复杂度指数当自然规律。结局后检验、未来依从性或未来药物调整不能进入基线预测；若研究动态风险，只在新的 landmark 更新，重新定义预测目标与窗口。

基线预测使用正则化回归、GAM、梯度提升，连续结局报告 MAE/RMSE、校准及区间；二分类报告区分、校准和罕见事件 PR 指标；时间到事件按竞争风险与观察覆盖设计报告固定时点校准、时间依赖性能。不能把治疗策略的条件风险预测差直接叫患者真实获益。

T4 因果问题与纯预测分开立项

若问题是“哪些患者更可能发生事件”，T2/T3 是预测；若问题是“对这个合格人群，启动 A 而非 B 会怎样”，必须另立目标试验式合同：资格标准、同时满足的时间零点、两种可执行策略、分配概念、随访、结局、因果对比和分析计划。

首选在同一治疗阶段、共同适应证下的主动对照新使用者/明确调整策略。示例估计量可以是目标人群在固定随访时点的平均连续疗效差或安全事件风险差；时间长度由问题及观察覆盖确定。启动策略效应与持续遵循策略的效应不同：前者需说明后续停药换药如何处理，后者需处理依从性、换药、死亡、失访和相应时间变化混杂。不能默默将停药者剔除，制造耐受者选择。

若联合方案允许启动后一段宽限期，不能把“后来成功加上第二药”者追溯为从第一天就暴露，否则容易引入不死时间。优先选择同时可判定策略的决策点；确需动态策略时，另评估克隆—删失—加权等设计及其数据要求，不为方法先进而强用。

以临床知识和因果图定义混杂结构，不自动把所有治疗后变量调整进去。可比较倾向评分加权/匹配、标准化结局回归及交叉拟合的双重稳健估计，报告权重分布、共同支持、平衡与有效样本量；选择性检测、病情严重度和医生偏好仍可能导致未测混杂。无共同支持就缩小目标人群或停止比较，而不是靠复杂模型外推。

获益异质性可在支持充分时探索预设交互、DR-learner 或因果森林，首先控制过拟合和多重探索。个体真实治疗效应不可同时观察，不能给“预测个体获益准确率”。应报告可识别的人群/亚组效应、不确定性，以及在明确假设下的独立集政策价值；模型建议的策略价值要用合适的离线估计或进一步前瞻研究评估，不能仅以 SHAP、特征重要性、AUROC 或回顾性推荐一致率证明因果有效。

两项可以直接讨论的默认研究配置

以下用具体疾病、策略与时间窗填满研究合同，作为优先试跑配置，而非要求导师另行提供设计。它们是可调整的研究示例，不是治疗建议、已获批方案或已验证最优选择。是否保留比较策略，取决于实际临床资格、共同支持、观察覆盖及与团队任务的契合；所有变更须在锁定测试前记录并冻结。

示范卡一 二型糖尿病加药后的应答与持续性

研究问题与人群。 默认选择成人 2 型糖尿病门诊患者，既有二甲双胍治疗记录至少 90 日，在同一加药阶段首次新增 DPP4 抑制剂或 SGLT2 抑制剂。既往 180 日未使用被比较的两类药，且源系统对该回溯期具有可核查覆盖。两种策略必须在病历所能确定的临床条件下均可考虑；心衰、慢性肾病等强选药适应证、肾功能限制、明确禁忌及医保或费用造成的决定性选药原因应识别并限制，不能仅凭药名认定两组可比；倾向评分共同支持只是必要检查，不能替代可交换性假设与未测混杂分析。该配置用于研究复杂联用与暴露质量的增量，不比较两类药在所有患者中的优劣。

时间零点与暴露。 index 为新增药首次处方日期，以处方启动策略定义主分析；若能取得发药/服药证据，建立相应敏感性暴露定义，不将首次处方等同首次服用。比较 DPP4 抑制剂加用与 SGLT2 抑制剂加用两个类别层策略，保留具体成分、剂量、同期药物变化和启动原因。若药物类内异质性或共同支持不足，则收窄至具体成分或符合共同支持的人群；若仍不足，不开展相对效应推断。

主要终点与观察窗。 默认疗效为 index 后第 90 日附近的 HbA1c：在第 60 至 120 日窗口内选最接近 90 日的测量，等距时取较早者；基线取 index 前 90 日内、决策时已可见的最近一次 HbA1c。该窗口测量是约 90 日应答的代理，不冒称所有人恰在第 90 日检测；记录实际测量日，做时点调整及更窄窗口敏感性分析。分析随访 HbA1c 并调整基线，变化值作为辅助描述。第 120 日前有无续方及停换原因另设持续性终点；处方中断仅称记录中的治疗持续性代理，不等于真实停药。未复诊者保留在流失描述和缺失分析中，不填正常或无事件。安全结局须单独定义；不能因 HbA1c 下降就声称总体净获益。

输入与两条分析路线。 特征仅用 index 前可见的年龄、基线 HbA1c、肾功能、共病、体重可得性、既有用药、剂量、近 180 日药物变更、联用负担及来源覆盖。预测输入还包括 index 时已经选定的启动策略、初始成分与剂量，或按策略分层建模；后续依从性和停换药可作过程/敏感性分析，不能作为启动时预测输入。先做预测路线：正则化回归和梯度提升，依次加入真实暴露代理、条件化联用证据；报告 MAE/RMSE、校准、患者聚类区间及各暴露质量层表现。再按 T4 另立比较路线：估计共同支持人群中两种启动策略在预设约 90 日测量窗口 HbA1c 上的平均差异，标准化或交叉拟合双重稳健估计，报告平衡、权重与有效样本量。失访、同期加药和未测选药原因做敏感性分析；不能把风险预测差冒充治疗效应。

执行与交付。 先输出资格与暴露 SQL、180 日回溯和随访窗口测试、患者流程与缺失图；再冻结患者分组及较晚时间测试集、完成基线与信息增量比较；最后交付应答与持续性数据字典、可复现分析包及带不确定性的分层结果。若随访 HbA1c 严重选择性缺失，先完成暴露/持续性方法研究，不把有随访者结果直接推广到所有加药患者。该默认配置不预先承诺样本数、效应方向或因果结论。

示范卡二 肺炎抗感染治疗四十八小时复评

研究问题与人群。 默认选择成人、非 ICU 病区、临床记录支持肺炎并启动静脉系统抗菌治疗的住院患者。以本次感染 episode 首次实际给药后 48 小时为 index，仅纳入此时仍住院、存活且处于待复评治疗阶段者。每个 episode 先取一个决策点，同一患者可有多次 episode 但不跨开发和最终测试组。结论针对到达 48 小时复评点的人群，不外推初始治疗期间所有患者，也不比较从入院起的因果疗效。

当前标签与未来终点分开。 当前任务是 index 时存在的可行动药理学问题，默认先聚焦剂量/给药方案与当时肾功能信息不匹配、重复覆盖或缺必要监测信息三类。两名药师依据冻结的本地规则、产品说明书与适用规范，独立判断需复核、未识别问题或信息不足，并对模型盲法；规则版本和裁定证据留存。未来客观预测任务则为 index 后 72 小时窗口内已观测肌酐峰值相对 index 前最近可见肌酐的变化量；基

线默认限前 24 小时内，不满足者标记基线不足。index 前已有肾功能异常、已有临床记录的肾损伤与此前稳定者分层描述，不能把既存状态写作新发事件，变化量也不自动表示新发或进展的诊断。这个连续实验室表型不是 AKI 诊断，更不是已归因于 DDI 的伤害；不据其单独判断抗菌疗效。

可见特征与观察边界。 纳入 index 前 24 小时可见的肌酐及变化、既往肾功能记录、体重可得性、当前抗菌药与合用药实际剂量/途径/给药间隔、既往抗菌暴露、病情与感染部位信息、此时已回报的培养/药敏和检验采集频率。48 小时之后才可见的培养结果、药师建议、停换药和出院信息不能进入输入。未来 72 小时缺肌酐不填零变化；死亡或出院导致的截断单列，报告观测覆盖、采样日和检测次数，不将有化验者视为全体无偏样本。

方法与比较。 当前问题识别依次比较冻结规则、药物数量/已知药对风险、临床变量的正则化模型、加真实给药、再加条件化证据与方案表示。若标签直接依据规则形成，应报告规则对标签的结构优势；学习任务的价值须落在来源整合、规则覆盖外经独立裁定的问题或复核排序，不能把复现规则称优于临床专家。连续肌酐变化用回归/梯度提升独立建模，比较是否加入实际暴露后有增量，同时分析检测策略造成的选择偏倚。当前标签报告各类 PPV/灵敏度、每百复评点告警及固定药师工时检出；连续结局报告误差、校准与预测区间，不跨任务拼一个准确率。

执行与交付。 先展示一条完整的 48 小时可见信息回放、实际给药 episode 构建和人工标签手册；再完成患者分组、较晚时间测试、训练内处理与独立校准，输出规则/临床/暴露/条件证据的阶梯比较和严重错误清单。最后交付可复用的肺炎复评队列定义、病例审核包、预测基线与拒答策略。若缺执行时间则降为处方复评研究；若肌酐观察不足则只保留当前药理学问题任务。任何结果均先供研究评价，不自动产生处方改变；欲比较抗菌方案疗效须另立可比策略及 T4 合同。

分割与泄漏审计是每个任务的组成部分

- 1. 患者分组：** 同一人的所有就诊、episode 及决策点进入同一开发/测试分组。跨院若无法合法可靠识别同一患者，说明残余重叠风险；不要仅按处方行随机分割。患者可在真实未来使用时重复出现，但“既有患者时间更新”与“全新患者泛化”须分开设测试任务。
- 2. 时间留出：** 按决策时间先开发后测试，并保证训练标签的随访成熟度，不让训练样本的结局延伸到未来测试期而造成知识可见性混淆。设置必要的训练截止/随访隔离区，记录临床流程和药物品种变化。测试集不能反复用于选特征、阈值、校准或终点。
- 3. 医院/病区留出：** 在条件允许时留出院区或中心，比较单纯冻结模型、仅截距/校准更新和正式再训练三种情境，校准更新须另留目标域校准样本，不能在最终测试结局上校准后再称未经更新外部验证。
- 4. 药物分组：** 药对 warm-start 评价已知药物的新配对；one-cold 为一药未见，both-cold 为两药未见。无向药对 A+B/B+A 统一；盐型、同成分不同产品和近重复产品按照目标任务成组，避免名义新药。若评价新结构再考虑骨架分组，但不能以结构分组代替患者外推。
- 5. 来源与研究家族：** 同一试验、专利、论文及其复述、同批实验的不同表行按家族分组；抽取测试不能混入其改写。条件推广要留出条件簇/制剂，而不是只随机留出同药对近重复记录。检索语料与模型参数版本冻结，记录大模型可能见过公共测试材料的限制。
- 6. 图与多任务泄漏：** 构图时去除最终测试标签及其等价/派生边；审查药物特征是否由目标标签库生成。可使用的无标签图结构要事先声明是归纳还是传导设置。多任务中另一任务的标签只有在部署时真实可得才可作为输入，不能通过共享药对把测试答案间接传回。
- 7. 处理管线：** 填补、归一化、特征选择、降维、文本词典调优、采样、阈值和超参仅在开发折内拟合。采用嵌套且按人/来源分组的交叉验证；训练、校准、最终测试用途分清。解释性分析也不能借测试结果反向改任务后继续称独立测试。

不把药物跨药对重复一概指控为泄漏。它可以是合法 warm 任务，但其结论只能支持相应任务。每个 release 输出 split manifest、去重规则、时间可见性规则及冻结哈希。

缺失 校准与不确定性

缺失机制按未做检测、做了未回报、来源未覆盖、无法链接、数值异常、原文未报告和不适用区分。预测模型可探索缺失指示和训练内填补，因缺失本身可能反映临床检测策略，但须在跨时间/跨院中压力测试；若性能依赖“医生已经怀疑所以加做检查”，应明确它的决策位置和可迁移风险。按预设时效使用前值，不无限向前填充肝肾功能；没有随访不填无事件。多重插补适合某些分析目标，但不能自动修复非随机失访或治疗影响下的检测。

小样本、罕见事件与标签不平衡时，不以过采样后的分布评价 PPV 或校准。比较类别权重、适度采样与原始分布基线，训练采样的概率偏移需处理；最终在真实目标患病/事件率的独立样本报告。生存结局的删失权重也需假设与敏感性分析。

分类模型报告校准曲线、截距/斜率、Brier 等适合任务的指标，不用一个分箱 ECE 证明所有亚组可靠。Platt/温度或保序回归等校准方法只在独立开发校准集选择；样本不足时不强用高自由度校准。连续结局报告预测值与观测值关系及区间覆盖。评估不同时间、中心、肾功能层、年龄与多药水平的校准，不把某亚组不显著解释为公平。

不确定性来源至少区分参数不稳定、标签不确定、条件缺失和分布外。模型集成、保形预测或距离/密度判别均可作为候选，但其保证依赖数据条件，时间/中心漂移时不能承诺无条件覆盖。拒答阈值由开发数据和可接受风险/复核容量预设，测试阶段锁定；拒答病例仍需统计结局和严重错误，避免把困难人群从评估中消失。

人工验证与统计评价

建立两条独立审核链：证据/暴露表型审核验证字段与时间重建；临床标签审核验证可行动问题或客观终点。评审材料、盲法、训练案例、分歧类别和第三人裁定规则预先写清。对可行动标签评审只给 index 前信息，对随访事件评审可给完整随访，但不暴露模型输出；同一评审不能先看模型建议再当独立金标准。

既做随机代表性样本，也对模型分歧、新药、稀有条件及高危错误过抽样。总体性能须使用代表样本或合适的抽样权重，困难集单独报告。报告逐类一致率、阳性/阴性一致性及适合标签结构的一致性统计；不能只用一个 kappa 掩盖低流行率和标签难度。药师未能达成稳定定义时，先修订任务，不强训一个看似高分模型。

性能区间按独立单位计算：患者任务按患者聚类 bootstrap，来源任务按研究家族，跨中心性能保留中心层不确定性；不能按药对表行/细胞/每日记录给虚窄区间。主要终点、比较和亚组先确定，探索项明确标记。模型选择不仅看显著性，还看增量大小、校准、严重错误、工作量与维护成本。

样本量由目标估计精度、事件率、有效自由度、缺失和聚类结构决定。先用获批可行性统计或可信先验进行区间/模拟规划，不写保证若干万患者，也不套一个固定“每变量若干事件”当万能门槛。样本不足时收窄标签和模型、报告宽区间或转为方法/可行性研究，不制造稳定亚组获益结论。

中药与组学的可选嵌套扩展

只有明确一种产品/组方版本且真实使用可追踪时，才在慢病包中开启“产品—暴露—共同治疗—表型—应答”扩展。继承团队已有临床—基础、质量与分子轨迹思路，不声称首次提出暴露、因果或分层。消渴方、消渴丸、玉泉丸及不同商品不能互替；产品、批次、实际组成/剂量、加减方和西药调整必须保留。

建模按临床基线、临床加真实用药、再加锁定 PANEL 三级比较。样本按患者和检测批次隔离，检测批次不能与病例/治疗组完全重合；批次校正与特征筛选在开发集内完成。只有治疗前样本可用于基线预测，治疗后组学可作反应/机制探索，但不能倒灌成为启动时预测特征；中介研究另需更强假设。

临床疗效为主要目标，分子状态“更接近健康”不能替代患者获益；血中检测到某成分不能证明靶组织作用。团队的胰岛单细胞或 REO 结果不可直接当外周血治疗反应模型。AB4/A3-6 现有证据是细胞/动物与靶点层级，不作为申请人近期自主人体试验主线。没有合规采样与检测合作时，先做产品/暴露和临床结局研究，删除完整分子机制闭环承诺。

可交付资产与推进门槛

第一阶段可以在不接触患者数据时启动。交付任务卡、标签手册、公开证据原型、规则/抽取基线、全合成时间线、split manifest 和泄漏审计清单；用现有真实来源演示字段抽取，不伪造模型结果。申请人可以现场展示 schema、提取脚本逻辑、错误分析和传统机器学习实验合同，体现实际技术能力。

第二阶段是获批院内可行性与方法基准。交付源字段可得性、医嘱与执行暴露的一致性、标签一致性、缺失/随访图以及冻结临床基线。能重建暴露和可观察结局才进入患者预测；只有医嘱时降级为处方过程研究；只有药理学裁定时不写客观获益研究。

第三阶段是锁定应用与外推。在一项主要抗感染或慢病任务上完成简单强基线、条件/暴露增量、时间留出、亚组校准和拒答评估。真实药物策略无重叠时不开展因果比较；模型无增量或工时成本高时，保留阴性结论及适用边界，而不是不断更换终点找显著。

第四阶段才讨论静默与前瞻验证。前瞻静默不向医生展示未验证建议，评估时间可见性、数据延迟和潜在工作量；真正提示、药师介入或改变治疗须另有授权与伦理/质量方案。维持现有药师与抗菌管理服务，比较对象、污染和实施依从性明确；提示被接受、药物减少、预测分数变好均不能自动认定临床获益。

公开发布字典、抽取/表型定义、代码、合成数据、评估协议与经审核聚合结果，真实患者数据默认院内可信环境运行。FAERS 发挥候选风险与证据优先级价值，自发报告的重复、选择性和未知分母决定其不能直接提供事件发生率、疗效对照或因果 DDI 真值。数据库、代码、模型及全文许可分别审查，公开可读不等于可再分发。

大数据挖掘与交互信号的发现及独立确认

本节跨资源 A/B 及 T1 至 T3 开展探索性发现，不属于纯 T0 文献抽取；独立确认后才考虑是否进入 T4 因果研究。

真实多药与时序模式挖掘

资源 B 的另一项独立方法任务，是从纵向数据发现“哪些组合或序列值得继续研究”，而不是把所有共现药对再装进一个分类器。可预设三类候选：同一治疗阶段的频繁药物集合、药物新增/停换与事件之间的顺序模式、至少三药的方案超边。先用 FP-growth/Eclat 等频繁项方法、PrefixSpan 等序列方法与可解释超图统计建立基线；只有数据规模、模式异质性和任务需要支持时才引入学习式图模型。申请人可以主导患者篮子、暴露区间、时间约束、机会分母及挖掘脚本，而非只向算法同事提供数据表。

篮子须由问题决定：每患者一次预设 episode、某个 landmark 前的已知暴露，或有记录支持的重叠给药集合；不能把整次住院出现过的药自动当成同时联用。先固定成分/制剂与剂量粒度，再做闭合/最大频繁项约简，保留临床有理由的非频繁组合。支持度至少同时报告独立患者数、符合资格 episode 数和暴露机会数；重复长期处方和频繁就诊不能无限增加支持。序列须有可解释的先后与最大间隔，事件后新增药不能倒推成导致事件的风险暴露。

“常见”和“异常共现”不同：lift、置信度等关联规则指标会受药物边际频率、药数、适应证与病区影响。比较保留边际用药率与患者药物负担的机会基线，必要时在适应证、严重程度、中心和时间层内置换或使用合适空模型；报告观察值与机会期望的差异，而不是看到高 lift 就称药理协同。超边模型要比较仅主效应、两两项及加入三药项的增量，判断第三药是否增加解释，不能把“含三药患者风险高”直接叫三药相互作用。

若把预设安全/疗效结局接到模式后，暴露形成窗口必须先于结局窗口，限制标签泄漏和事件诱发治疗的反向因果；有潜伏期/既存事件时先定义风险集和滞后敏感性分析。数据驱动挖出的模式是候选假设，不是因果发现。可结合稀疏回归、层级收缩与经验贝叶斯稳定性分析控制稀有计数，但模型收缩不能弥补缺乏实际暴露或结局分母。

探索阶段预先定义药物空间、最高阶数、最小患者支持、主要结局族、时间范围及合适的多重检验/FDR策略。阈值按运算规模、罕见风险价值和统计精度确定，不按哪个结果好看倒推。使用患者聚类重抽样考察模式稳定度；在独立后续时间或中心按冻结模式、同一暴露/结局定义验证效应方向、区间和增量，不再重新挖掘后称验证。验证时也处理预先规定的多重性；回顾病历确认候选是否只是“某病更重所以多用药”的标志，再决定是否值得进入独立 T4 比较研究。

负结果同样有价值：若高阶模式几乎均被药物数量、疾病阶段或两两项解释，就报告复杂度收益不足；若低支持导致区间极宽，保留为待验证证据缺口，不输出“高危黑名单”；若现有规则已覆盖多数高价值模式，研究贡献可转向可复用的挖掘基准、暴露测量与误分类影响，不为制造新意强加 GNN 或更高阶组合。

FAERS 交互信号与医院确认的双来源路径

FAERS 是申请人既有方法积累的延伸，但不作为医院队列的替代品。冻结季度/下载版本，保存原报告 ID、病例/随访标识及来源版本，先按官方字段含义处理同病例后续报告、跨期重复和明显重复；去重规则、保留版本、日期冲突与不确定重复分别留痕。还须规范商品名/成分/复方、给药途径可得性、事件术语版本、报告时间与发生时间可得性、药物角色及共同报告药。不能把同名/相近年龄报告臆合为同一个患者，也不把 FAERS 与医院记录做患者级拼接。

首轮可比较单药信号、药对报告比值类信号、加入交互项的报告层模型，以及对稀有组合收缩的层级/经验贝叶斯方法。主要可疑、次要可疑和伴随用药角色按明确规则处理，并作角色定义敏感性分析；两药同被报告不代表已证实同时暴露或被同等怀疑。药对交互项相对选定报告模型的偏离、加性/乘性报告尺度的差别，要逐项说明，不能直接命名为药理协同或临床相互作用效应。高阶组合仅在报告支持与稳定性够用开展，并与单药/药对模型比较必要性。

所有分母是相应数据库内符合定义的报告，不是用药人数或人时。报告量变化、适应证、严重程度、上市时间、监管/媒体关注、共报告偏倚和信息缺失均影响信号。采用药物/事件族分层、最低报告支持、收缩、多重性控制及前后时间留出评估稳定性；日期不全时不可据报告提交时间推定真实暴露—事件顺序。外部重复检出仍是信号复现，不是发生率或因果证实。

医院确认路径依次为：候选组合和事件的药学合理性审阅；查询真实给药与事件可观察范围；盲法病历裁定时间关系及其他原因；在有资格风险集、可比对照和数据质量时建立独立队列；只有识别条件成立才进入 T4 因果估计。FAERS 与院内结果不一致应分析角色编码、暴露误差、人群和事件定义差异，不选择性只报一致信号。无院内可验证病例时，结果仍可作为可追溯假设资源，但不升级为临床警报或用药禁忌。

第四部分 启动交付与证据核验边界

分阶段交付与进入条件

阶段	我可以主导的工作	可展示成果	进入下一阶段的依据
设计与公开原型	选主题边界、schema/字典、证据卡、提取规则、合成队列、Q1-Q5 查询合同	可审查的数据资源蓝图、真实来源小样例、全合成患者流、许可清单	导师能看出科学问题与方法接口；不需要患者数据启动
主题证据集 v0	系统检索、去重、抽取、双人标注、条件/证据域冻结	带溯源主题数据集与独立测试集、提取误差和工时报告	可复核质量与许可足够；若关键条件大面积缺失，改做证据缺口研究而不伪造完备库
获批院内可行性	源系统字段核查、身份/事件关联、少量合规人工核对、ETL 与质量报告	院内 CDM 小范围实现、字段可得性与缺失图、episode 准确性	执行时序/观察覆盖满足所选问题；只有医嘱则限制任务，不假称实际暴露
冻结研究队列	版本化抽取、预设表型、独立裁定、时间/机构分层	可复现数据包、公开字典/合成示例、资源/方法分析	事件数、可比性、标注一致性支持应用研究；样本量由实际盘点与目标精度确定
临床/机制应用扩展	联用证据桥、慢病/抗感染任务、必要中药产品子模块	与算法及疾病研究同事共同的研究任务和论文	不越过伦理、数据权属与实验能力条件；先证明价值再扩规模

可用一个**设计性工作量例子**向导师展示可执行性：公开原型先以 20–30 个成分、若干本院相关制剂类别、约 100 条跨域证据卡试跑字典/审校流程；这不是已完成数量，也不是正式资源规模或足够统计效力的承诺。随后依据每域有效原始证据、人工成本和研究问题重新定量。医院队列不提前写“十万患者已可用”。

申请人与协作方的研究责任

- **申请人学术牵头**：确定研究对象与标签本体、四域边界、条件/时序规则；提出测量误差与证据不一致问题；设计金标准与质量评价；主导结果解释与论文。
- **申请人工程主导**：设计表和主键、编写/组织可复现 ETL 与抽取流程、词典映射和质量规则、版本管理、构建研究队列和数据发布包；个人实现与协作实现的边界，在任务分工与版本记录中明确。
- **申请人方法主导与协作**：主导数据抽取、实体链接、弱监督和数据挖掘的任务定义与基线，与团队共同推进条件化表示及任务建模。
- **团队合作价值**：既有分子/多任务方法、资源建设、组学与机制知识帮助完成可迁移方法和跨域解释；不是团队只出模型、申请人只出病例。
- **院内与外部协作**：信息部门控制身份与权限；护理/静配核实实际接触；临床专家参与事件裁定；统计学合作者审查估计目标；需新兼容实验/药代/组学时由有条件平台承担。资源支持、审批与实验能力分别确认。

可冻结的发布包与质量门槛

建议每次 release 提供：README 范围与用途、data_dictionary、schema、source_registry 与许可矩阵、vocabulary_mapping、provenance、annotation_manual、quality_report、split_manifest、cohort_definitions、synthetic_examples、queries、CHANGELOG、checksums。真实医院数据表不放入公共发布包。

硬门槛不是任意“总体准确率 95%”：

- 主外键完整、来源与版本可追溯、许可状态明确、敏感数据披露审核通过，是发布前必须满足的条件；不合格记录隔离而非混入。
- 关键危险字段（否定、年龄边界、给药方式、剂量/单位）只要抽查发现系统性错误，暂停相关批次发布并定位规则；更正后重新抽检。
- 抽取/映射性能目标在试标后按用途和可接受误差预设，并报告置信区间；不能先跑结果再挑一个门槛过关。
- 覆盖率、未知率和错误率同时报告；提高准确率而大量拒收困难记录不是无条件进步。
- 原始来源变更、纠错或撤回触发影响分析，给受影响的研究 release 出勘误，不覆盖旧结果而失去复现能力。

证据核验范围

本计划以申请人提供文献及学校、研究群、原始出版来源和官方资源说明为依据，进行有边界的定向核验，不声称完成全领域系统综述或每篇全部补充资料复现。下表解释主要结论所依据的阅读深度，使后续工作可按优先级补足证据，而非把不确定内容填成既成事实。

证据组	已核内容	使用边界
DeepIDC 和 MDDC	本地原文方法 结果 表格及讨论；MDDC 正文 791 至 796 页	支持药对标签与任务定义；未重跑代码 不作实现层 泄漏指控
ACRM	PubMed 作者与摘要 出版商公开预览	完整正文和补充方法缺失；不推定有效性细节及原 模型可复现性
糖尿病与冠心病研究	核心方法 研究对象及疾病标签；2023 冠心病 采用出版社材料	分类不等于治疗效应；未核全部补充数据
P450Rdb TCM2COVID scRiskCell	原始论文与资源说明的定向核对	可支持资源类型和边界；下载许可及当前可运行性 需分别审查
中药 2021 2024 2026 框架	本地思路稿相关章节与图表	支持研究思路承接；不是已完成临床试验或数据库
DDInter Stabilis OMOP	官方资源介绍与规范	已有能力作为竞争基线；未逐条审核第三方所有数 据
DailyMed 说明书 示例	指定 setID 说明书相关警告段落及版本日 期	仅该来源声明；不能代替中国具体制剂标签或实验 记录
INSPIRE TriMaster 治疗选 择与配伍竞争研究	原刊摘要和相关设计段落；各研究核验深 度不同	用于排除过度首创表述；e-Harvis 和部分治疗比较 仅摘要级，不推定完整算法
FDA PhysioNet MIMIC	官方公共数据局限 权限与使用条款说明	不表示已获资格 伦理审批 数据访问或外发许可

52 行原始文献清单是申请人收集的候选记录，不是 52 篇已逐字精读的独立论文。其中包含重复和非论文模型记录；部分 PDF 结构损坏。名为“2021 Antibiotic combinations”的本地 PDF 实际上是 Information Fusion 的糖尿病研究，DOI 为 10.1016/j.inffus.2021.02.015，不能用于证明 ACRM 方法。本计划不以文件名代替原文身份，不以缺损文本填补未经核定的细节。

进入实际研究前优先补足的四项证据

1. 合法获取 ACRM 完整正文和补充资料，确认可比任务、标签与数据接口；若不能取得，则独立定义患者级任务并明确比较范围。
2. 对选定主题的第三方数据逐项核定访问、分析、派生发布和全文再分发权限，完成原始证据追溯和竞争资源头对头比较。
3. 获院内授权后盘点给药、检验可见时间、院外观察及患者链接字段，依据质量与事件数决定可实施的研究问题，而非先承诺样本规模。
4. 若选择中药或新增兼容性实验，明确具体产品、检测条件、样本及有能力合作平台；观察研究结果不自行授权临床干预。

可复核的结论边界

本计划已经形成的是研究设计、技术任务和证据地图；尚未形成新的数据库发布、临床模型性能、治疗效果或导师合作承诺。论文路径是潜在成果结构，是否独立成文取决于实际科学贡献与数据质量，不预定论文篇数或期刊。即使复杂模型没有超过规则，仍应如实报告测量误差、适用范围和证据不足，但不把所有阴性结果都自动包装成高价值创新。