

医疗数据治理初探及 高水平数据集建设实践

宣武医院从临床数据治理到高水平数据供给的探索

临床数据



临床治理



临床应用

presentation
framework

汇报框架

01

从“数据汇聚”到“数据可用”

P2-P6 为什么医院需要重新认识医疗数据治理

02

“高水平医疗数据集”是什么？

P7-P14 不是病例数量越多越好

03

宣武医院实践

P15-P22 从一个颈动脉支架数据集开始

04

信息科怎么复制

P23-P29 高水平数据集建设关键点

05

从一个数据集走向医院数据能力

P30-P32





首都医科大学宣武医院神经外科

Department of Neurology and Neurosurgery, Capital Medical University

中国国际神经科学研究所

China International Neurological Research Institute (CINRI)

01

第一部分

从“数据汇聚”到“数据可用”

2026年HIT专家网讲师讲坛课件
版权归演讲者所有

系统很多，真正用数据时仍然很难



科研、AI与专病需求仍然遇到

- 01 找不到
- 02 找到但不完整
- 03 系统口径不一致
- 04 非结构化程度高
- 05 诊疗过程串不起来
- 06 每个项目重新整理

数据汇聚，不等于数据可用。

第一部分 | 从“数据汇聚”到“数据可用”

医院数据建设正在进入第三阶段

认知
升级



回答：有没有数据

回答：数据在哪里

回答：数据能否解决问题

需求已经从“导数”升级为“生产数据产品”

过去：单次数据导出

临床：帮我导一批数据

交付结果：表格或数据库导出

工作方式：一次一清洗、一次一解释

现在：围绕场景生产数据产品

科研：构建可研究的患者队列

AI：影像 + 报告 + 病灶标注

药械研发：治疗过程 + 随访 + 疗效 + 结局

交付结果：可复用、可评价、可持续更新的数据集



信息科面对的已不是单纯导出需求，而是数据产品生产需求。

治理的本质是围绕场景持续加工数据

采集



关联



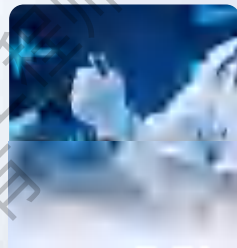
标准化



结构化



临床语义



质控



标注



应用



持续加工分散数据，使其可理解、可计算、可信、可复用。



医院内临床数据治理不是再建一个平台，也不是再给非结构化的数据做一遍抽取。

医疗数据治理最终要回答三个问题

01

数据有没有？

01

数据对不对？

01

能否解决问题？

治理的终点不是“拼凑数量”，是“能解决问题的数据”。



首都医科大学宣武医院神经外科

Department of Neurology and Neurosurgery, Capital Medical University

中国国际神经科学研究所

China International Neurological Research Institute (CINRI)

02

第二部分

“高水平医疗数据集”是什么

2026年HIT专家网
版权归演讲者所有
演讲论坛课件

第二部分 | “高水平医疗数据集”是什么

高水平数据集必须同时满足六个条件

判定
标准



数据水平不由病例数量单独决定

10万例，也可能是低水平数据

只有截图

只有阳性 / 阴性

报告无法对应

来源不可追溯

标签不清、授权边界不明

1000例，也可能是高水平数据

数据链条完整

临床信息充分

图文稳定对应

精准标注

来源可追溯、质量可评价

病例数量决定规模； 数据颗粒度和可用性决定水平。

第二部分 | “高水平医疗数据集”是什么

原始
数据

原始数据质量首先要保证“真”

CT / MRI

优先保留原始 DICOM

支持序列、参数和空间信息追溯



数字病理

优先保留 WSI

避免缩略图丢失诊断级细节



超声 / 内镜

优先原始视频或高质量文件

保留过程信息与关键帧上下文



截图、缩略图或不可追溯文件，不能替代诊断级原始数据。

临床文本要从“结论”升级为“临床语义”

低质量表达

影像数据
阳性

高水平表达



仅有“阳性/阴性”通常不足以支撑高质量训练，诊断依据同样重要。

第二部分 | “高水平医疗数据集”是什么

多模态数据的关键是“稳定对得上”

关联
质量

第二部分 | “高水平医疗数据集”是什么

不同任务决定不同的数据粒度

任务
驱动

任务 分 类	最低要求 病例级标签	高水平要求 多标签、病种、严重程度
报告生成	影像 + 报告	所见、结论、病灶绑定
检测	BBox	3D BBox + 病灶属性
分割	Mask	3D Mask + 类别字典
VLM	图文配对	病例—病灶多层绑定
VQA	问题&答案	答案可追溯至诊断依据

数据质量：没有脱离应用场景的绝对高质量。

正式交付的是一套“数据产品包”



第二部分 | “高水平医疗数据集”是什么

建立分级后，医院才能回答“有哪些好数据”





首都医科大学宣武医院神经外科

Department of Neurology, Capital Medical University

中国国际神经科学研究所

China International Neurology Research Institute (CINRI)

03

第三部分

宣武医院实践

- 从一个颈动脉支架数据集开始

第三部分 | 宣武医院实践

宣武实践不是从“大平台”开始的

起点是一个具体问题

颈动脉支架手术数据，
如何形成可研究、可流通的专
病数据集？

项目过程中才暴露出



数据



治理



审批



交付



匿名化

使用
追踪

先解决真实场景，再把项目能力沉淀为医院级能力。

颈动脉支架专病数据来自多个业务系统



专病数据集不是单系统数据，而是患者全过程的多源关联。



患者主索引



就诊关联



手术关联



检查关联



时间序列



随访关联

最终形成



患者诊疗旅程

真正困难的治理发生在“字段→医学变量”



治理不是改字段名，而是把临床问题转化为可计算变量。

IT规则

结构与一致性

- ★ 手术日期不能早于入院日期
- ★ 出院日期不能早于手术日期
- ★ 年龄与出生日期应一致

临床规则

医学合理性

- ★ 用药与诊断路径符合临床逻辑
- ★ 特定治疗仅出现在适用患者
- ★ 诊断—治疗—结局相互印证



专业规则引擎示例 复杂医疗指标提取

维度展开与聚合

案例

出血前用药（41个选项）

原始数据

- ▶ [“阿司匹林”，“氯吡格雷”，“PPI”]
- ▶ 展开为 41 列独立变量：阿司匹林=**是**，氯吡格雷=**是**，华法林=**否**...
- ▶ 聚合为 2 列药理分类：抗血小板药物=**是**，抗凝药物=**否**

数据质量规则必须由信息科与临床专家共同定义。



源头奠定数据治理基础，临床专家规则、结构化模板和系统质控，专病库内均是高质量标注数据。





信息中心牵头，临床专家定义规则，多部门协同保障数据质量、安全和应用转化

共建项目工作组

信息科（数据管理部门）

成果转化办（数据应用）

业务科室”至少三级联动工作组，必要时“信息科（数据管理部门）

成果转化办（数据应用）

业务科室+影像/病理/检验科

数据服务/标注团队



对数据流程及流转的审核



对数据应用场景使用的监管



对数据交付的把关



01 需求分析

数据需求与变量定义



02 数据评估

源数据质量评估与预处理



03 规则构建

临床专家知识图谱使用



04 批量处理

大规模数据智能结构化处理



05 质量提高

数据质量验证与优化



06 数据安全自查

权限设置检查
数据安全自查
数据质量六性自查

流程
SOP
标准化

一次专病项目最终沉淀为医院级闭环



交付不是终点；使用过程和价值反馈会反向更新治理规则。



首都医科大学宣武医院神经外科

Department of Neurology, Capital Medical University

中国国际神经科学研究所

China International Neurology Research Institute (CINRI)

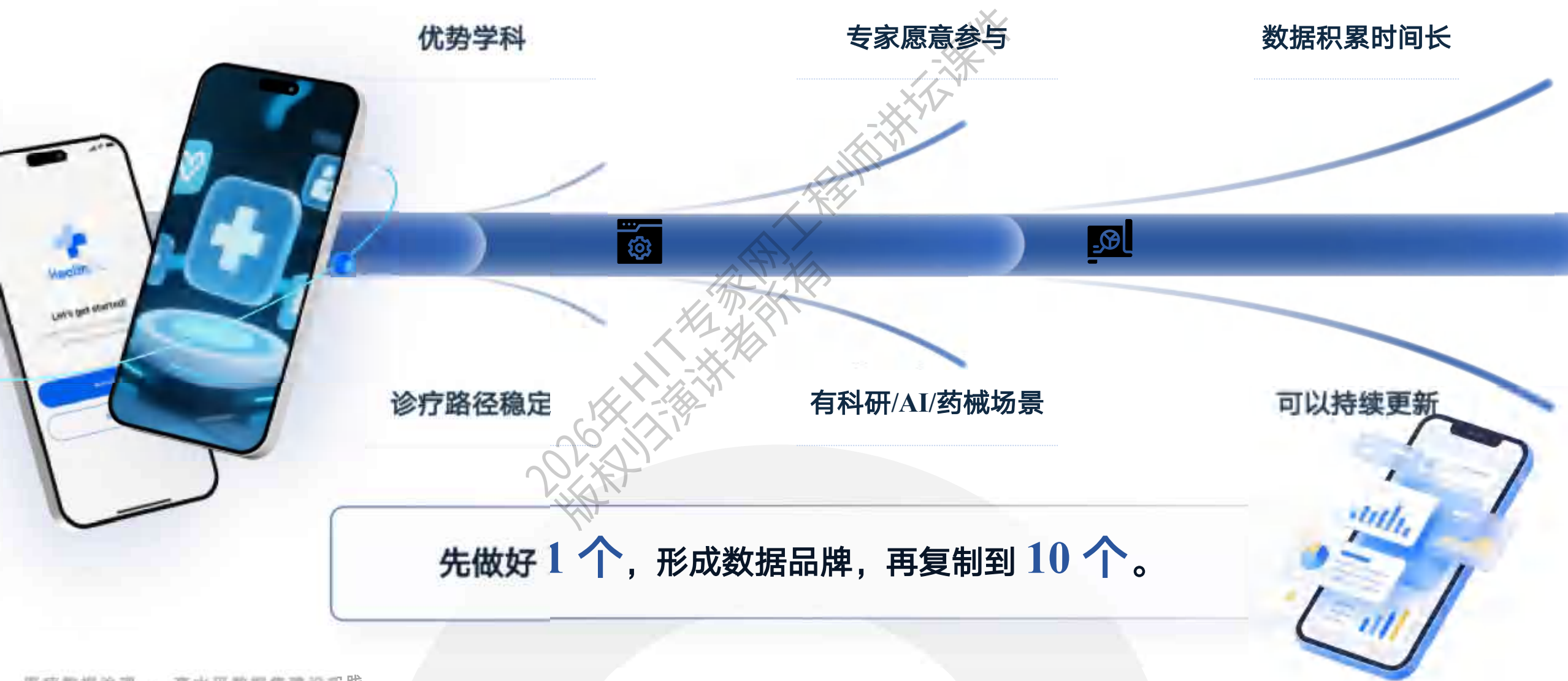
04

第四部分

高水平数据集建设关键点

■ 高水平数据集建设关键点

复制的起点不是全院，而是优势专病



先确定用途，再决定治理范围



科研



AI训练



RWE



药械研发

用途共同决定



样本体量



时间跨度



数据粒度



字段范围



标注形式



质量要求

没有清晰用途，治理范围就会无限扩张。

专病数据模型要还原患者全过程



专病患者 360° 纵向数据模型



安全与合规

最小必要

仅采集与诊疗或供给场景直接相关的字段

标准治理

对齐医疗数据标准，解决格式和口径问题

脱敏预处理

按场景选择策略，并保留使用价值

可追溯

来源、加工、审批、交付和使用全程留痕

01

临床专家

Clinical expert

- 变量定义
- 标注规则
- 医学质控
- 金标准确认



02

信息科

Information Office

- 数据抽取
- 治理加工
- 系统实现
- 安全红线



03

体系与流程

System and Process

- 伦理与合规
- 院内审核
- 授权备案
- 成果与版本管理



高水平数据集不是信息科单独做出来的，而是多角色共同生产。



高水平数据集不是一次验收出来的，而是在使用中不断迭代出来的。

可复制的不是数据库，而是生产机制

容易复制，但价值有限



真正应复制的能力

复制表结构

复制接口

复制一次性交付



复制场景选择方法

复制规则与质控体系

复制跨部门协同和闭环



让下一专病建设更快、更稳、更低成本，才算形成医院能力。



首都医科大学宣武医院神经外科

Department of Neurology and Neurosurgery, Capital Medical University

中国国际神经科学研究所

China International Neurological Research Institute (CINRI)

05

第五部分

从一个数据集走向医院数据能力

2026年HIT专家网-讲师讲坛课件
版权归演讲者所有

数据资源目录

知道有哪些数据、在哪里、由谁负责

数据标准规则

统一语义、编码和字段口径

数据质量规则

把IT校验与临床逻辑固化

专病模板

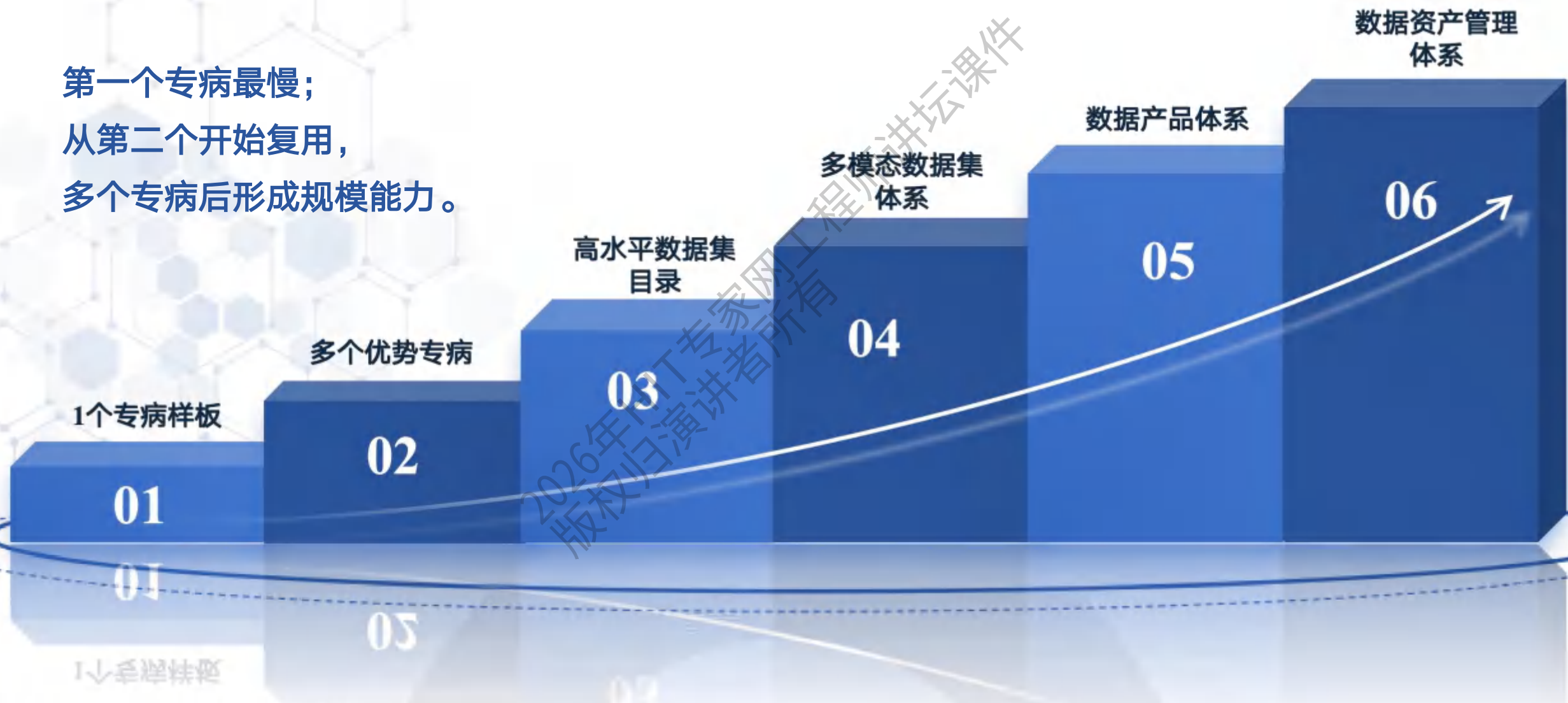
复用患者模型、变量和标注框架

匿名化与安全策略

按场景复用合规供给能力

单专病样板最终要升级为医院级能力

第一个专病最慢；
从第二个开始复用，
多个专病后形成规模能力。



把每一次项目沉淀成医院能力

◆ 数据汇聚只是开始，数据可用才是目标

◆ 高水平数据集一定是场景驱动

◆ 最难的是让临床知识进入数据

◆ 信息科与临床必须共同生产

◆ 一次项目要沉淀为可复用能力

2026年HIT专家网工程师讲坛课件
版权归演讲者所有