解决方案

英特尔® SGX
联邦学习
智慧医疗

医渡云借多方安全计算技术打破医疗科研数据孤岛

医渡云
YIDUCLOUD

“临床医学离不开真实世界的研究，而真实世界研究依赖高质量数据。我们正通过构建更为安全和高效的多方安全计算解决方案，让更多高质量数据被充分利用，成为推动医疗科研事业高速发展的助力。为合法合规地打破因数据安全要求引发的‘数据孤岛’问题，我们与英特尔一起，结合其 SGX 技术构建了联邦学习方法所需的硬件可信执行环境，让不同医疗机构的数据协同实现‘更安全+更高效’的双重优势。”

闫峻博士
首席 AI 科学家
医渡云

前言概述

算法的演进、算力的提升、数据的持续扩展，是当今医学领域人工智能（Artificial Intelligence，AI）应用落地与发展，特别是在精准诊疗与医疗科研中开展实践的重要基石。这三者中，数据尤为关键，其价值不仅仅在于体量，更重在维度和来源，因此医疗科研所需的高质量 AI 模型构建，通常都离不开多方和多维数据的协同参与。

然而，此前该领域一直被数据隐私保护和信息风险防控要求所制约，各方数据多以数据孤岛的型态存在，多方数据协同很难实现。为了更好地挖掘多方和多维数据中的巨大价值，同时更好地兼顾到其隐私和安全的防护，中国医疗智能行业头部企业医渡科技旗下医渡云开始着手研发更为安全和高效的多方安全计算解决方案，包括与英特尔开展深入合作，利用英特尔® 软件防护扩展（Intel® Software Guard Extensions，以下简称英特尔® SGX）技术的优势，为新方案导入基于硬件可信执行环境（Trusted Execution Environment，TEE）的联邦学习方法，为医疗科研中参与多方计算的敏感数据和代码提供更为可靠的安全防护。

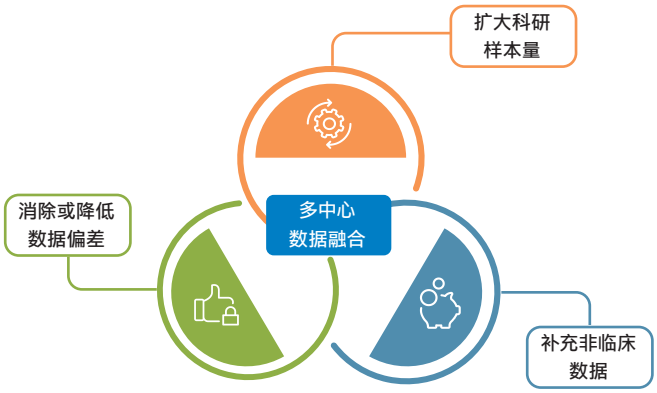
背景与挑战：医疗科研亟需更好的数据融合与价值挖掘

“AI+医疗”这一融合正成为医疗技术加速数字化、智能化演进的重要推进器，沙利文研究报告指出，我国医疗智能行业市场规模预计将在 2030 年超过 1.1 万亿元人民币[1]。与其他行业和领域 AI 落地的条件和规律相似，除算法和算力的进步外，海量、多维的数据也是其持续、快速发展的基石，而一系列医疗信息化系统的完善以及数字化医疗设备的普及，让医疗大数据早就达到了厚积薄发的状态，换言之，医疗科研领域构建 AI 模型，加速研发与探索的核心条件已经成熟。

不过，医疗科研毕竟是一个细分化的、复杂的、系统化的领域，尽管各个医疗科研机构自身都有大量的数据资产，但在体量和维度上仍有较大的差别，这对科研效率会有实质性的影响。毕竟，数据集的体量越大、维度越丰富，能够从中发现和学习到的特征就越多，基于此构建的 AI 模型的性能及应用价值也就越高。大量统计数据已表明，多中心研究机构的医疗科研效率往往会优于单中心机构，关键就在于多中心机构能借助多方数据的融合与协作，在数

据体量及维度上实现更大优势，进而也能对数据中的价值进行更为深入和全面的挖掘和利用。因此，医疗科研机构普遍期望能开展多方及多样化的数据协作。如图一所示，多中心数据融合可为医疗科研带来以下关键优势：

- **消除或降低数据偏差：** 研究区域以及方法、方式的差异，会带来不同研究中心间的数据差异，通过数据融合，能消除或降低数据偏差，使研究成果泛化能力更强；
- **扩大科研样本量：** 数据融合能够让不同研究中心间的临床数据得以共享，扩大科研所需的数据样本量，提升最终 AI 模型的性能；
- **补充非临床数据：** 许多长期跟踪的医疗科研数据还需要对社区医疗、家庭医生、体检机构以及可穿戴设备的数据实施融合。



图一 多中心数据融合带来的医疗科研优势

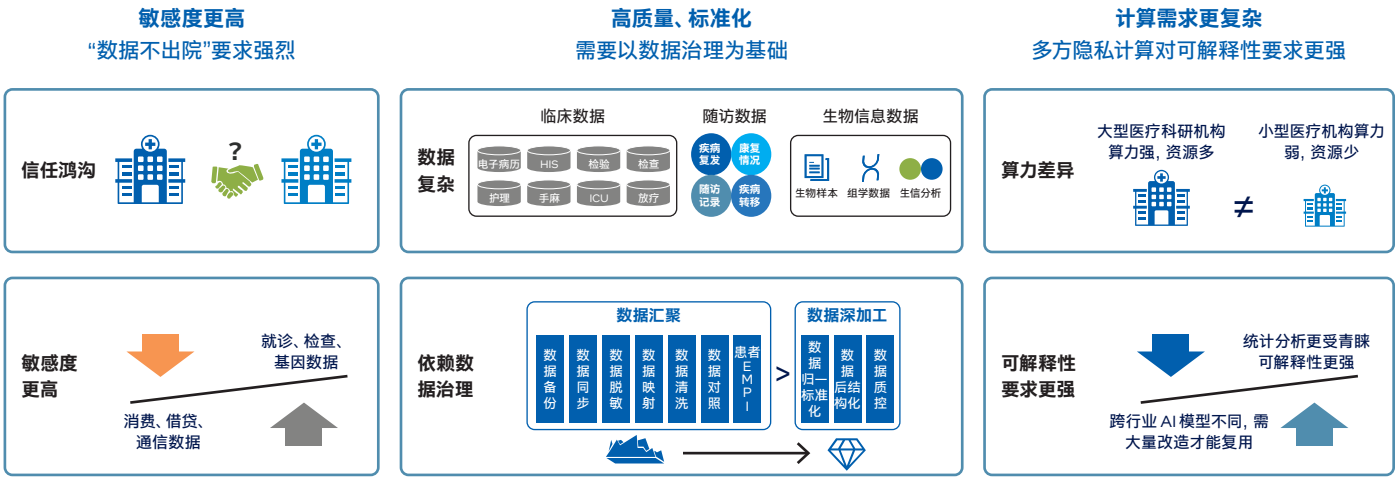
虽然多方数据协同好处多多，但在实践中这种融合和协同带来的数据安全问题也越来越受关注，在国家政策层面，中国已出台《个人信息保护法》、《数据安全法》等一系列法律法规来对数据安全和隐私信息予以保护。为此，医疗科研机构采取了一系列方法来规避风险，包括采用长链条的数据协同审批流程，以人工方式进行数据录入、转录等。但这些方法不仅耗时长、效率低，还缺乏质控且难以溯源，带来了严重的数据孤岛问题。

在这种矛盾的现实状况下，如何合法合规地解决数据孤岛问题，让医疗数据在融合的同时也能满足隐私保护和科研应用的双重需求，以及有望兼顾这两个需求的多方隐私计算技术，就成为了众多医疗科研机构关注的焦点。

不同于其它领域，医疗科研对基于多方隐私计算技术的数据融合有一些特定的需求，如图二所示，这些需求涉及：

- **数据敏感度：** 医疗科研场景下的数据敏感度很高，“医疗数据不出院”的需求非常强烈，因而在参与各方之间建立信任也非常困难；
- **数据融合标准化：** 医疗科研对数据的高质量要求，使之非常依赖数据治理。各个进行中的研究项目可能需要反复的调整纳排条件后，再进行全局性的安全聚合计算；
- **计算需求：** 医疗科研基于多方隐私计算技术的 AI 建模通常有着明显的行业特点，因此 AI 建模时对计算性能也有很高的要求。

为帮助众多医疗科研机构打造兼顾高效和安全需求的多方隐私计算能力，为医疗和健康行业提供更优的数据融合与数据科研价值挖掘能力，多年来一直深耕医疗 AI 与大数据技术创新的医渡云，



图二 医疗科研领域数据融合需求的特征

以强大的医学数据治理能力为后盾，通过自研 YiduManda 安全计算引擎为数据融合提供了联邦学习、联合统计、联盟区块链等核心技术保障。

这其中，采用 TEE 方案的联邦学习方法凭其在数据“可用而不可见”方面的独到优势，在各医疗科研机构的实践中收获了良好效果。与其他多方隐私计算方案相比，采用 TEE 方案的联邦学习方法具有以下优势：

- 医疗数据不脱离本地，各参与方可利用自身拥有的数据训练全局模型；
- 每个医疗科研参与方都可参与训练过程，模型损失可控；
- 训练过程能更好地兼顾隐私和安全需求，各参与方能在不暴露数据及加密形态的前提下进行联合建模。

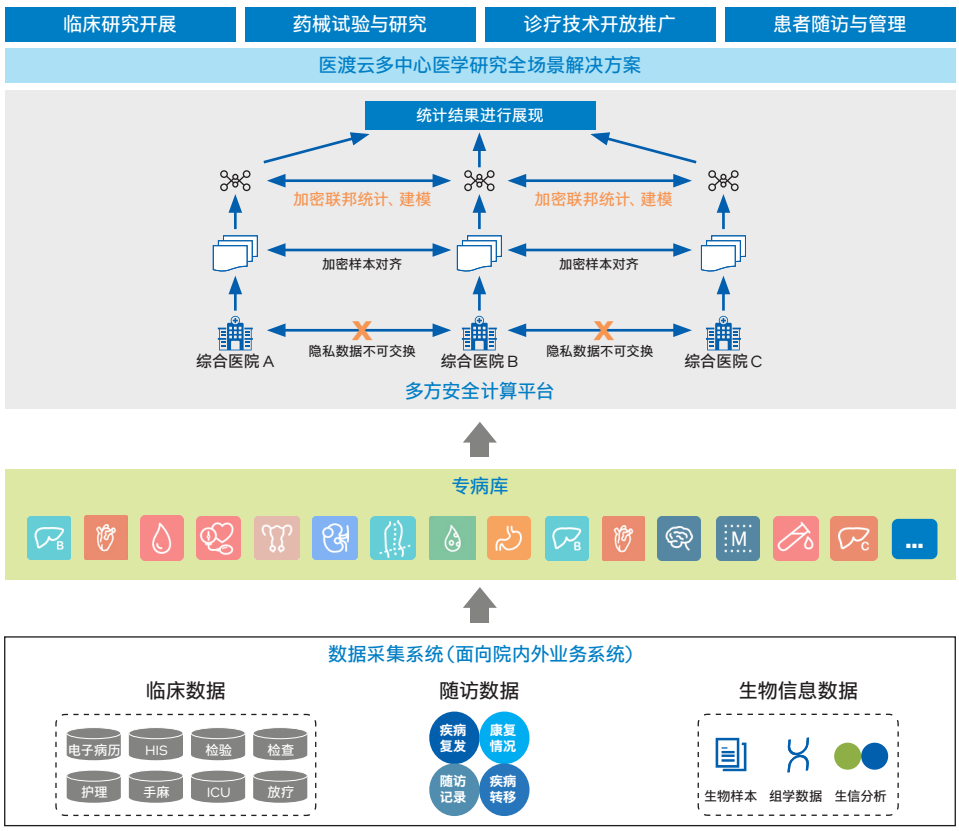
为此，医渡云正与英特尔携手，导入英特尔® SGX 来构建基于硬件可信执行环境的联邦学习方法，来为各医疗科研机构打造高效的多方安全计算解决方案。

基于英特尔® SGX，以联邦学习方法构建高效多方安全计算解决方案

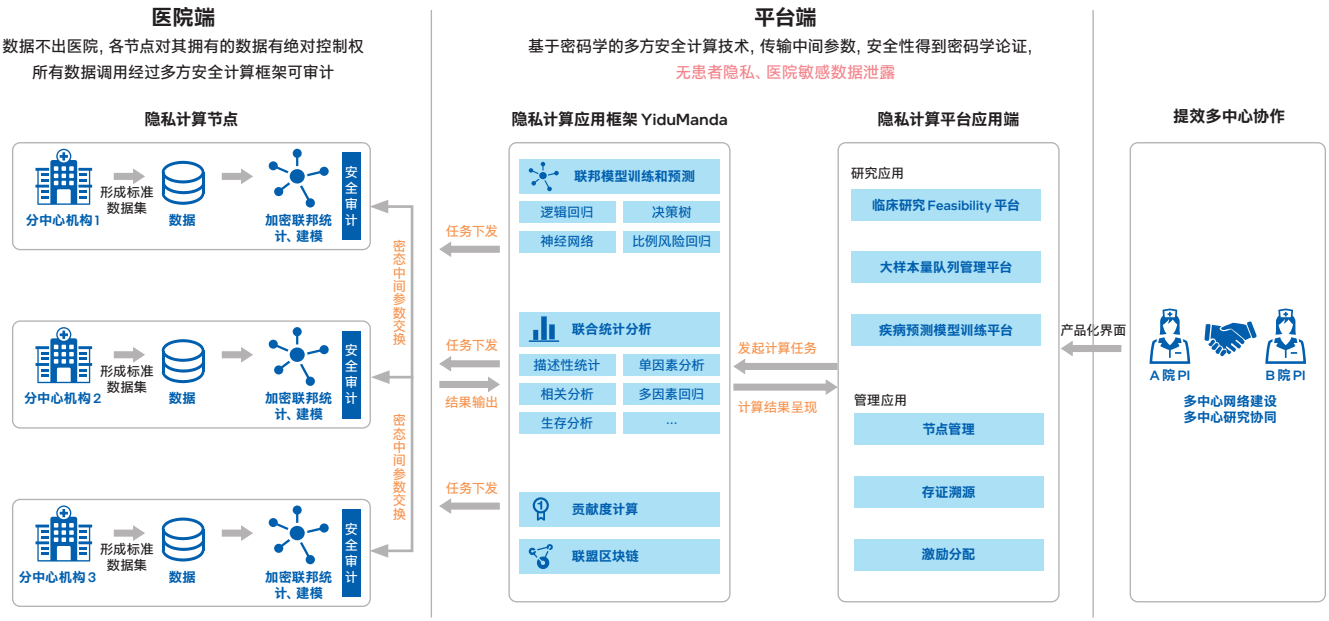
医渡云基于联邦学习等隐私计算方法打造的多方安全计算解决方案，其功能层面如图三所示，自下而上分别是面向院内外业务系统的数据采集系统、进行数据加工治理的专病库以及开展多方隐私计算的安全计算平台。在安全计算平台之上，医渡云又通过多中心医学研究全场景解决方案，部署了一系列面向多样化医疗科研场景所需的上层应用能力，如临床研究开展、药械试验与研究、诊疗技术开放推广、患者随访与管理等。

具体来说，方案中各层的功能和作用分别为：

- **数据采集系统：** 医疗科研机构开展临床研究所需的数据一般来自于研究机构的临床数据中心、随访中心、生物样本库以及生物信息中心；
- **专病库：** 采集后的数据需要执行同步、脱敏、映射等数据汇聚过程以及归一标准化、结构化等数据深加工过程。完备的数据加工治理流程，能帮助医疗科研机构按照研究目标，把各个科研参



图三 医渡云多方安全计算解决方案整体架构



图四 医渡云多方安全计算解决方案中医院端和平台端的协作模式

与方(医院或医疗机构)的多元异构数据转换成统一格式的高质量数据，通过数据抽取后建立满足研究所需的专病数据库；

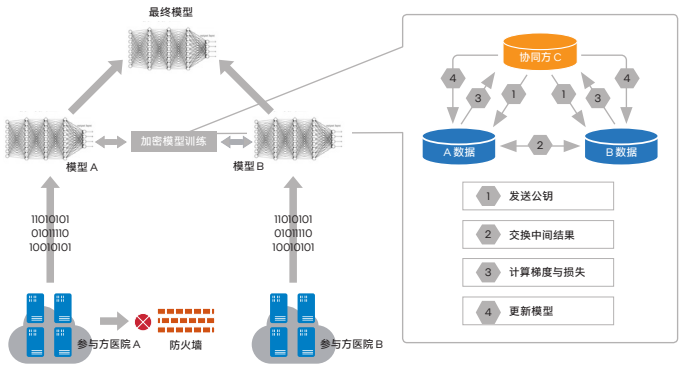
- **多方安全计算平台：** 医渡云自研的 YiduManda 以多方安全计算、联邦学习为基础，同时结合英特尔® SGX 将来自各个科研参与方（医院）的原始数据，通过联合统计、特性工程（Feature Engining）、逻辑回归（Logistic Regression，LR）、XGBoost 等方法进行联合统计分析和模型训练，并最终得到医疗科研 AI 模型以及相关深度学习模型。

在架构设计上，医渡云的方案采用了分布式的设计，可分为平台端（调度节点）和医院端（计算节点），其中：

- **平台端（调度节点）：** 部署在互联网数据中心或机构联盟的主中心私有云环境中，包括一套用于联邦学习等隐私计算的调度层框架以及相应的科研应用平台。应用层框架对各医院端隐私计算节点进行统一的管理和协调，并对多方安全计算的任务进行统一调度；

- **医院端（计算节点）：** 部署在医院的私有云环境中，通过隐私计算节点间的协作，能保证数据在不出医院的前提下完成联邦学习等多方隐私计算过程，各个节点对其所有的数据有绝对控制权，所有数据调用经过多方安全计算框架可审计。

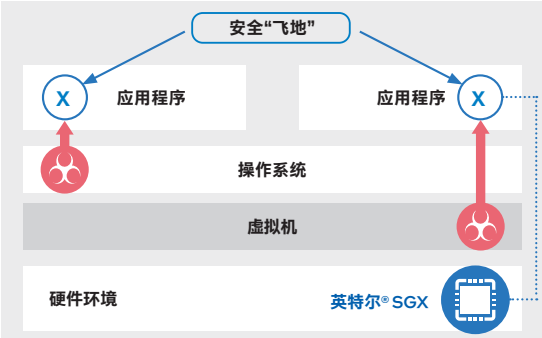
基于上述功能与架构设计，各医疗科研机构之间开展基于联邦学习的模型协同训练的过程如图五所示，数据准备阶段是在各个参与协同训练的医院或医疗机构本地完成的，准备好的数据通过程序接口加载到医院端中，随后平台端会调度完成模型的协同训练过程。参与训练的医院端通过加密信道与其它参与方完成通信和计算，并最后完成模型的优化训练。



图五 基于联邦学习的模型训练

在方案的具体部署中，医渡云引入了英特尔® SGX 来构建基于 TEE 的联邦学习方法所需的硬件可信环境。英特尔® SGX 能在内存的特定硬件环境中构造出一个可信的安全“飞地”(Enclave)，为医疗科研过程中参与多方计算的敏感数据和代码提供更强的安全防护。

如图六所示，与其它技术方案相比，英特尔® SGX 一方面为敏感数据与程序构建了隔离的硬件环境，使安全保护机制独立于软件应用、操作系统或硬件配置之外，从而令保密性和完整性大幅提升；另一方面，独立的“飞地”设置可让关键的应用程序和数据更有效地避开来自硬件驱动程序、虚拟机乃至操作系统的攻击，带来更强的安全性。基于英特尔® SGX 提供的这些优势，各医疗科研机构就可将数据分析、模型训练及推理所涉及的数据运行在“飞地”中，通过访问控制为这些应用代码和数据提供更可信赖的安全保障。



图六 英特尔® SGX 技术实际作用示意图

在性能表现上，英特尔® SGX 基于硬件层面的安全保护机制，可使敏感数据与应用程序获得来自基于英特尔® 架构的处理器强劲性能的加速或助推，从而更好地解决方案中性能和安全的平衡问题，在某些对计算性能和安全等级要求都很高的医疗科研场景中输出更为全面的应用优势。

2021年发布的面向单路和双路服务器的第三代英特尔® 至强® 可扩展处理器，已集成了英特尔® SGX，并为此提供了更优的支持，其高端型号最高可在双路系统中支持 1TB 容量的保留加密内存区域（Enclave Page Cache，EPC），这对于医疗科研机构进一步扩展 AI 模型训练与推理的数据规模至关重要，因此该处理器在医渡云多方安全计算解决方案中也扮演了关键角色。当然，除了集成 SGX 技术外，该处理器对内核微架构、I/O、内存性能及容量的改进和提升，及其内置的英特尔® 高级矢量扩展 512（英特尔® AVX-512）和英特尔® 深度学习加速（英特尔® DL Boost）技术对 AI 应用的硬件加速能力，也为方案涉及的复杂计算需求提供了有力支撑。

落地及展望

得益于服务全国 800 多家医疗机构，覆盖 60 个疾病领域所积累的深厚经验，医渡云可以为面向医疗科研领域的多方安全计算解决方案带来专业的方案设计[2]，而第三代英特尔® 至强® 可扩展处理器不仅为方案带来了数据处理所需的强劲算力，其内置的英特尔® SGX 也为方案提供了更可靠的数据安全防护，使用英特尔® SGX 构建的基于 TEE 的联邦学习方法，为多方数据安全、高效的联合建模打造了更为可信的硬件环境。在面向医疗科研领域的实践中，医渡云已经支持多家医院和医疗科研机构开展了一系列基于多方安全计算解决方案的联合研究项目。

综合以上优势，医渡云目前已通过了中国信息通信研究院在隐私保护计算技术上的两项认证，分别为《基于多方安全计算的数据流通产品技术要求与测试方法》与《基于联邦学习的数据流通产品技术要求与测试方法》[3]。

随着采用英特尔® SGX 的多方安全计算解决方案得到越来越多的客户认可，医渡云已计划将该方案作为未来核心产品的一个基础组件来提供默认的隐私计算能力，并根据用户需求提供服务。

面向未来，医渡云也将继续携手英特尔，针对多方安全计算中的多中心临床研究解决方案开展更为深入的合作，这些合作包括：将英特尔® SGX 及相关技术和框架用作其整体隐私计算解决方案中的重要选项，借助该技术在安全特性和性能上的双重优势，为那些对计算性能要求较高的场景提供更优的支持，并在单中心内部的隐私保护、跨中心联邦学习等更多场景中探索英特尔® SGX 的运用。当然，这些合作的目标都是一致的，即为医疗科研事业的发展提供源源不断的技术助力和数据积累。

intel XEON

[1] 如欲了解更多详情，请访问：http://www.frostchina.com/?p=17404
[2] 数据来源于医渡云，如欲了解更多详情，请访问：https://www.yiducloud.com.cn/
[3] 以上二项标准属于由中国信息通信研究院牵头起草的隐私计算系列标准，详情请咨询中国信息通信研究院相关信息

扫一扫，英特尔医疗解决方案一键直达