数据来源与可靠性 · 一页纸简报

BIORI BIOTECH | 宝锐生物 营销中心 | 守护生命健康，缔造美好生活

2026-10
内部说明 · 可复核
怎么取数 · 怎么不瞎编

4 层	130 篇	521 项	10 份	0 条
数据来源分层	本地开放全文语料	工作台数据 JSON	单次取证全文核验	含 2%/3% 限度的条款

一 · 功能 能做什么

1. **文献调研 → 证据表**
   三源检索（Europe PMC / OpenAlex / Crossref）→ 取开放全文 → 逐篇抽「定量+条件」→ 引用机械校验；每篇标 ✓已读全文 / ⚠仅摘要
2. **法规 / 标准取证**
   先取官方附件原文（.doc/.docx/PDF/扫描件 OCR），全文逐句扫描，输出「全文字数 × 关键词命中数」并落盘全文，结论可复算
3. **行业 / 竞品调研**
   官网、财报、标准、专利入口；先查实测可达矩阵再写抓取，不拿「打不开」当「没有」
4. **销售数据统计与勾稽**
   ERP → JSON → 自包含 HTML/Excel；金额现算，与源文件透视表/进度表交叉校验（差 0.00）
5. **竞品与市场情报**
   竞品官网/财报/注册证/招投标定时扫描；先查可达矩阵再写抓取，抓不到就标「未接入」，不拿「打不开」当「没有」
6. **对客话术支撑**
   把技术结论转成客户可用的证据卡：明确区分「法定要求」与「企业内控」，不让销售在客户现场把内控说成法规
7. **交付形态**
   自包含 HTML（零外链）/ PDF / Excel；附来源清单与缺口清单，任何人可复核
8. **知识沉淀与复用**
   结论存为事实卡 / 证据卡入知识库，工作台看板同步；下次同类问题直接调卡，不重复检索

二 · 来源 数据从哪来（四层）

L1 宝锐一手数据 优先
ERP 订单原始导出 | 数据层 **521 项** JSON（客户/价格/商机/客诉/活动/KPI） | 飞书活动·商机·客诉（MCP 实查） | 产品与公司资料 | IMA 两个知识库

L2 本机已审证据语料 先查再联网
开放获取全文 **130 篇**（DOI 命名、可逐句检索） | 调研稿 **4 篇** | 已审事实卡（数值三件套 + 使用边界）

L3 法规 / 标准 / 期刊原文 唯一可作依据
NMPA · CMDE 官方附件（.doc/.docx 直链） | 地方药监 | YY/T 标准（扫描件 OCR） | 期刊 PDF | WHO / EMA / WOAH / PMC | 专利

L4 公开检索接口 免 key
Europe PMC（首选，带摘要/被引/PMCID） | OpenAlex（引用分位替代 JCR 分区、含预印本） | Crossref（核卷期） | PubMed | Semantic Scholar（引文雪球）

不假装有的 边界
Scopus / Exa / NotebookLM 未接入 | JCR 分区抓不到（403）用 OpenAlex 分位 | 付费墙只给缺口清单 | 站方限流如实写覆盖率

检索纪律 两条
① **先本地后联网**：本机语料（全文 130 篇 / 事实卡 / 数据 JSON）先查，查不到再检索外部；② **中文词先翻英文**：英文库对中文词会「有结果但全错」（实测返回完全无关文献），必先做术语映射

口径纪律 唯一权威源
同一指标只认一个源：年度目标＝**目标基础数据_2026.json**；营收金额＝ERP；集团归属＝CRM。多处出现不一致时以权威源为准并回溯订正

三 · 可靠性 怎么保证不瞎编

1. **原文主义** 结论只来自原文全量；搜索摘要、二手文章、镜像转载不得作依据
2. **量化留痕** 每份文件留「来源直链 + 全文字数 + 命中数 + 原句」，全文落盘
3. **假阴性三关** 编码（GB2312）· 解析（doc/docx/扫描件）· 覆盖（页数）逐关自检
4. **覆盖不足不下结论** 拿不全就写「6/11 页」+ 缺哪页 + 为什么缺，不把「没查到」说成「没有」
5. **多方互证** OCR 结论 × 标准简介/条款框架 × 等同采用关系（=ISO）；ERP 用源表透视表校验
6. **来源分级** 一手（原文/官方附件/期刊全文）> 二手（行业网/媒体）> 厂商口径（标不可作依据）
7. **数字现算** 金额/占比一律脚本现算、可勾稽；无口径的数写缺口，不写「约 X 万」
8. **纠错留痕** 错就公开写进交付物（本次 3 处），不悄悄改
9. **权限与合规** 客户数据只取授权源、不外发、不进外部 AI；无权客户不确认不否认

交付前必过 4 道验收

- □ 检索式 / 纳排 / 起止年写进报告，可复现
- □ 每篇标 ✓已读全文 / ⚠仅摘要，全文覆盖数对得上
- □ 引用校验通过：0 个未知 id、0 条空证据
- □ 缺口清单逐条列清（标题 / DOI / 为什么非它不可）

四 · 别人最常问的四句 我们的回答

数据是不是编的?
每条结论都挂在原文/原表上；没有来源的断言标 [unverified]，绝不用相近文献冒充。

为什么跟网上说法不一样?
网上多为二手转述。本次实测：某二手文章把「冻干复溶稳定性」记到另一份指导原则名下，核官方附件后**撤回该引用**。

我能自己复核吗?
原件、提取全文、**全文字数 × 命中数**、扫描日志全部留存，可按路径逐字比对。

查错了怎么办?
错就公开写进交付物。本次 **3 处自查纠错**：误引撤回、GB2312 假零命中、编码判定启发式。

工具链 `lit_search.py` 检索/全文/雪球 | `sources.py` 引用校验 | `epmc_probe.py` 精确查 | `scan_docs.py` 全文扫描·字数×命中数

实证 · 冻干产品「水分 均值≤2% / 单瓶≤3%」依据核查 已结案

10 份	11.8 万字	1 处	3 处	1 条
法规/标准/综述 全部取原文	全文逐句扫描 非摘要、非二手	「水分」二字 出现次数（仅设备名/方法）	自查纠错 误引撤回·假零命中·编码	结论：法定体系无此条款 该数值属**企业内控**

BIORI BIOTECH · 宝锐生物 | 电话 0756-8699969 | www.biori.com | marketing@biori.com | 珠海市香洲区水岸一路88号