Meta-Analysis Project Documentation: Association of Metformin Use and Cancer Incidence

Generated on: 2025-10-24 07:40:52 **Creator:** krisztian.sugar@frogs.hu ("budapest" team)

---

1. Input and Scope Definition

1.1. Research Topic

The objective of this meta-analysis project was to systematically investigate the association between **metformin use and cancer incidence**. This topic encompasses both the potential chemopreventive effects of metformin in non-cancer populations and its prognostic impact in patients with pre-existing malignancies.

1.2. Data Source Limitation

Due to licensing constraints, the systematic search and subsequent full-text acquisition were restricted exclusively to publicly available open-access articles accessible via the PubMed API and associated DOI links.

---

2. Methodology: Literature Search and Pre-filtering

2.1. Database Search Strategy

The literature search was executed solely using the PubMed API. A specialized Large Language Model (LLM) was employed to generate a comprehensive set of seven distinct search queries designed to maximize sensitivity across various aspects of metformin, diabetes, and oncology research, while explicitly excluding secondary literature types (systematic reviews, meta-analyses, and narrative reviews).

Search Queries Generated by LLM:

1. `(metformin OR Glucophage OR biguanide OR dimethylbiguanide) AND (cancer OR neoplasm OR carcinoma OR malignancy OR oncogenesis OR incidence OR risk reduction OR tumor) NOT (systematic review[pt] OR meta-analysis[pt] OR review[pt])`
2. `('Metformin'[MeSH] AND ('Neoplasms'[MeSH] OR 'Cancer Incidence'[MeSH]) AND ('Cohort Studies'[MeSH] OR 'Case-Control Studies'[MeSH])) NOT (systematic review[pt] OR meta-analysis[pt] OR review[pt])`
3. `(metformin[tiab] AND (cancer incidence[tiab] OR neoplasm risk[tiab])) NOT (systematic review[pt] OR meta-analysis[pt] OR review[pt])`
4. `(metformin OR Glucophage) AND (cancer OR neoplasm) AND (cohort study[pt] OR case-control study[pt] OR longitudinal studies[mesh]) NOT (systematic review[pt] OR meta-analysis[pt] OR review[pt])`
5. `(metformin OR biguanide) AND (glucose metabolism OR insulin resistance OR AMPK pathway) AND (cancer OR tumor OR malignancy) NOT (systematic review[pt] OR meta-analysis[pt] OR review[pt])`
6. `(metformin AND (cancer OR neoplasm) AND incidence) AND (2014:2024[dp]) NOT (systematic review[pt] OR meta-analysis[pt] OR review[pt])`
7. `(metformin AND cancer incidence) AND (randomized controlled trial[pt] OR clinical trial[pt]) NOT (systematic review[pt] OR meta-analysis[pt] OR review[pt])`

Initial Retrieval Result: A total of **513** articles were retrieved from the PubMed database.

2.2. Abstract-Based Pre-filtering

The retrieved articles underwent an initial screening phase based on their abstracts and metadata using LLM analysis guided by predefined inclusion and exclusion criteria.

Inclusion Criteria (GOOD CANDIDATES):

- Clear randomized controlled trial (RCT) or systematic review methodology.
- Well-defined study population and intervention.
- Measurable primary and secondary outcomes.
- Statistical analysis reporting effect sizes, confidence intervals, or p-values.
- Clinical relevance and significance.
- Adequate sample size.
- Clear inclusion/exclusion criteria.

Exclusion Criteria (BAD CANDIDATES):

- Case reports or case series (sample size $n<10$).
- Editorial comments, letters, or opinions.
- Animal studies or *in vitro* studies only.
- Lack of control groups.
- Unclear methodology or outcomes.
- Preliminary or pilot studies without sufficient power.
- Studies with major methodological flaws.
- Conference abstracts without full methodology.

Sample Abstract Classifications:

PMID	Classification	Confidence Score	Reasons for Classification
39560490	Good Candidate	0.95	"['Retrospective analysis of two clinical cohorts (human NSCLC patients) with clear clinical outcomes (RFS, PFS, OS).', 'Utilizes defined statistical methods (Hazard Ratio, Confidence Interval) suitable for meta-analysis.']"
35378172	Bad Candidate	1.0	"['Preclinical study utilizing murine (mouse) models exclusively.', 'Focuses on mechanistic outcomes (T-cell function, transcriptomic analysis) rather than human clinical endpoints (OS, PFS).']"

Pre-filtering Result: 242 articles remained after abstract filtering.

2.3. Full-Text Acquisition

Full-text articles were downloaded using the PubMed API, with a fallback mechanism utilizing the DOI link for open-access content.

Download Result: 178 articles were successfully downloaded.

---

3. Methodology: Full-Text Classification and Data Extraction

3.1. Full-Text Classification Categories

The remaining 178 full-text articles were subjected to detailed LLM analysis to extract and categorize methodological and contextual information. The primary classification categories included:

- `article_type`: General manuscript format (e.g., Original Research, Guideline).
- `candidate_meta_analysis`: Suitability for quantitative synthesis (CANDIDATE/ NOTACANDIDATE).
- `cochrane_bias`: Risk of bias assessment (detailed in Section 4).
- `data_type`: Type of data presented (e.g., clinical, administrative, molecular).
- `species`: Species studied (e.g., *Homo sapiens*, *Mus musculus*).
- `study_type`: Research design (e.g., RCT, Cohort Study).
- `clinical_test`: Specific measurements or assays performed.
- `cohort`: Detailed characteristics and size of participant groups.

3.2. Full-Text Candidacy Assessment

The `candidate_meta_analysis` classifier determined the final pool of articles suitable for quantitative data extraction.

Sample Candidacy Classifications:

PMID	Classification	Confidence	Assessment Rationale
37225730	CANDIDATE	High	"The study meets all primary assessment criteria. It is a retrospective cohort study with defined comparative groups (VFI tertiles, metformin use) and provides robust quantitative data (HRs with 95% CIs) for survival outcomes (RFS and OS). It is strongly suitable for inclusion in a quantitative meta-analysis."
40105683	NOT_A_CANDIDATE	High	"The document is a clinical practice guideline/consensus report (secondary literature) and therefore fails the primary criterion requiring original quantitative research. It is unsuitable for inclusion in a quantitative meta-analytic synthesis."

Final Candidacy Result: 115 articles were confirmed as candidates for meta-analysis after full-text review.

3.3. Study Type Classification

The LLM categorized the research design of the candidate articles:

Sample Study Type Results (Randomized Controlled Trial):

PMID	Classification	Confidence	Evidence
36581893	Randomized Controlled Trial	High	"The study is a pooled analysis of data derived from two separate clinical trials: the Reach for Health Study (RFH) and the MetBreCS trial. The Methods section explicitly states that participants were 'randomly assigned to metformin... or placebo administration' and describes both source trials as 'randomized, double-blind, placebo-controlled trials.' The core methodology involves an interventional design where participants were prospectively allocated to an intervention group (metformin) or a comparator group (placebo) using random assignment, which aligns precisely with the definition of a Randomized Controlled Trial."
34629300	Randomized Controlled Trial	High	"The study is explicitly identified as a 'Randomized Phase II Study' in the title and the methods section. The core methodological characteristic is the random assignment of eligible, non-diabetic mCRPC patients into two groups: the intervention group (Docetaxel + Metformin) and the control group (Docetaxel + Placebo). Patients were 'randomly assigned at a 1:1 ratio.' The study followed participants prospectively from the point of randomization to assess outcomes like PSA response rate, Progression-Free Survival (PFS), and Overall Survival (OS), fulfilling all criteria for a Randomized Controlled Trial (RCT)."

3.4. Meta-Analysis Target Selection

Based on the comprehensive classification of available cohorts and clinical tests across the 115 candidate articles, a single target outcome was selected for quantitative meta-analysis due to resource constraints.

Selected Target Outcome:

`json { "selected_clinical_test": "Hemoglobin A1c (HbA1c)", "justification": "HbA1c is a standardized, widely reported laboratory test for long-term glycemic control, making it highly comparable across studies. It is clinically relevant for assessing the metabolic effects of interventions in diabetes, PCOS, and cancer-related metabolic dysregulation, all of which are heavily represented in the provided data.", "recommended_cohorts": ["Type 2 Diabetes Patients on Metformin", "Polycystic Ovary Syndrome (PCOS) Patients on Metformin", "Cancer Patients on Metformin"] }`

3.5. Comprehensive List of Extracted Clinical Tests and Cohorts

The full-text analysis identified a wide range of clinical tests and cohorts, reflecting the broad scope of metformin research in oncology and metabolic health. A partial list of identified clinical tests includes:

- **Metabolic/Endocrine:** Hemoglobin A1c (HbA1c), Fasting Plasma Glucose (FPG), Oral Glucose Tolerance Test (OGTT), HOMA-IR, Serum Insulin Levels, C-peptide Level, Total Testosterone (T) Assay, SHBG Assay, LH/FSH Assays, DHEAS Assay, IGF-1.
- **Oncology/Pathology:** Core Needle Biopsy, Estrogen Receptor (ER) Status IHC, HER-2/neu Status IHC, Gleason Score, AJCC Cancer Staging, RECIST criteria, PSA Test, Tumor Grade, Histological Confirmation of HCC/PCa/BC.
- **Anthropometric/Risk Factors:** Body Mass Index (BMI), Waist Circumference, Waist-to-Hip Ratio (WHR), Charlson Comorbidity Index (CCI), Ferriman–Gallwey Score (FGS), Smoking Status Assessment.
- **Imaging/Procedures:** Transvaginal Ultrasound Scanning, CT Scan, MRI, Mammography, Colonoscopy, Prostate Biopsy.

The identified cohorts span diverse populations, including:

- Type 2 Diabetes Patients (T2DM) (e.g., 23137378, 27026681, 32532851)
- Polycystic Ovary Syndrome (PCOS) Patients (e.g., 16764619, 29482528, 34726324)
- Colorectal Cancer (CRC) Patients (e.g., 27496094, 30084749, 32196659)
- Breast Cancer (BC) Survivors (e.g., 29788487, 33516778, 36581893)
- Hepatocellular Carcinoma (HCC) Patients (e.g., 32212089, 33661912, 36008432)
- Prostate Cancer (PCa) Patients (e.g., 27789181, 32035002, 36178848)
- Lung Cancer (LC) Patients (e.g., 26973204, 32532664, 33262518)
- Solid Organ Transplant (SOT) Recipients (e.g., 32159875)

3.6. Data Point Extraction

The multimodal Pro LLM processed the full texts to extract quantitative data points relevant to the selected outcome (HbA1c). The extraction focused on mean values, standard deviations, sample sizes, and effect measures for intervention and control groups